FINDING HIDDEN PATTERNS OF
HOSPITAL INFECTIONS ON THE NEWBORN IN TURKEY:
A DATA MINING APPROACH

İNCİ AKSOY

BOĞAZİÇİ UNIVERSITY

2010

FINDING HIDDEN PATTERNS OF
HOSPITAL INFECTIONS ON THE NEWBORN IN TURKEY:
A DATA MINING APPROACH

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts
in
Management Information Systems

by
İnci Aksoy

Boğaziçi University
2010

Finding Hidden Patterns of

Hospital Infections on The Newborn in Turkey:

A Data Mining Approach

The thesis of İnci Aksoy
has been approved by

Assist. Prof. Bertan Badur
(Thesis Advisor)

Assist. Prof. Sona Mardikyan
(Thesis Advisor)

Assoc. Prof. Osman Nuri Darcan

Assoc. Prof. Eyüp Çetin

Assoc. Prof. Ünsal Tekir

April 2010

Thesis Abstract

İnci Aksoy, "Finding Hidden Patterns of Hospital Infections on The Newborn in Turkey: A Data Mining Approach"

The increasing number of hospital infections with considerable morbidity, mortality and economic burden attracts the attention of not only the health-care environment, but also the whole society. In some medical centers, hospital infections are traced with more controlled and extensive surveillance methods, which adopt data mining applications. Data mining methods are applied to find the outbreaks, which cannot be determined easily by infection control teams.

This study presents an application of data mining methods for hospital infection detection in a newborn intensive care unit. The data set is provided by Department of Clinical Microbiology and Infectious Diseases, Eskişehir Osmangazi University, Faculty of Medicine. Decision tree, neural network and logistic regression classification models are built with holdout sampling and cross validation. In model comparison, accuracy and sensitivity measures are taken into consideration. Bagging and boosting methods are applied on neural network and decision trees in order to increase the performance of these models.

According to the results, antibiotics and urinary catheter usage, peripheral catheter duration, enteral and total parenteral nutrition durations, and birth weight for gestational age are prominent risk factors. Among the models, neural network performs well on hospital infections detection representing 83% accuracy and 30% sensitivity on test data set. Furthermore, the sensitivity is improved with boosted neural network model to 44%.

Tez Özeti

İnci Aksoy, "Türkiye'deki Yenidoğan Hastane Enfeksiyonlarının Gizli Örüntülerini Bulmak: Bir Veri Madenciliği Yaklaşımı"

Her geçen gün görülme sıklığı artan hastane enfeksiyonları, önemli derecede morbidite, mortalite ve ekonomik yüklere neden olmakta ve yalnızca sağlık sektörünü değil, tüm toplumu ilgilendirmektedir. Bazı sağlık merkezlerinde hastane enfeksiyonları veri madenciliği uygulamalarını kapsayan daha kontrollü ve kapsamlı sürveyans yöntemleriyle takip edilebilmektedir. Bu uygulamalarda veri madenciliği yöntemleri kolaylıkla belirlenemeyen salgınların tespitinde kullanılmaktadır.

Bu çalışmada, yenidoğan yoğun bakım ünitesindeki hastane enfeksiyonlarının tespit edilmesi için veri madenciliği yöntemlerinin uygulaması sunulmaktadır. Veri seti Eskişehir Osmangazi Üniversitesi, Tıp Fakültesi, Klinik Mikrobiyoloji ve Enfeksiyon Hastalıkları Bölümü tarafından hazırlanmıştır. Basit ve çapraz doğrulama yöntemleri ile karar ağaçları, yapay sinir ağları ve lojistik regresyon sınıflandırma modelleri kurulmuştur. Model karşılaştırmada doğruluk ve duyarlılık oranları dikkate alınmıştır. Doğruluk oranını arttırmak amacıyla karar ağaçları ve yapay sinir ağları modellerinde bagging ve boosting yöntemleri uygulanmıştır.

Sonuçlara göre, antibiyotik ve üriner kateter kullanımı, periferik kateter kullanım süresi, enteral ve total parenteral beslenme süreleri ve doğum ağırlığının gestasyonel yaşa oranı önemli risk faktörleri arasındadır. Yapay sinir ağları, test setinde %83 doğruluk ve %30 duyarlılık oranı ile hastane enfeksiyonları tespitinde başarılı olmuştur. Bunun yanında boosting yöntemi yapay sinir ağlarında duyarlılık oranının %44'e çıkmasını sağlamıştır.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisors, Assist. Prof. Bertan Badur and Assist. Prof. Sona Mardikyan, for their invaluable guidance, support, encouragement and motivation during my thesis.

I would like to thank Assoc. Prof. Osman Nuri Darcan, Assoc. Prof. Eyüp Çetin and Assoc. Prof. Ünsal Tekir for their participation in my thesis jury.

I would very much like to thank Prof. Dr. İlhan Özgüneş, Department of Clinical Microbiology and Infectious Diseases in Eskişehir Osmangazi University Faculty of Medicine, who shared the data set and answered my questions with patience.

I am deeply grateful to Kemal Özden and Devrim Melek Tunç for their invaluable support and encouragement.

I would also like to thank my director Ivan Cavinato, my manager Galip Berker and my colleagues Burcu Daş, Zeynep Helvacı, Seray Göklü and others in Yapı Kredi Bank for their precious support and comments, without their motivation and sympathy I could not finish my study.

I would like to share my sincere thanks to my cousin Atilla Onrat and my friends Esra Aceleoğlu, Olcay Kurnaz, Burçin Aksoy and Buket Bako. This thesis would not have been possible without their support and encouragement.

Finally, I would like to thank my family, my father Sezai, my mother Leyla and my sister Beril who were the inspiration and motivation in this study.

To the loving memory of my father Sezai Aksoy

CONTENTS

FIGURES

TABLES

CHAPTER 1

INTRODUCTION

Hospital infections are health care related infections, which indicate an important problem in both developing and developed countries because of high morbidity, mortality and heavy economic burden they cause. They can spread to public from patients, hospital staff and caregivers. Bacteria, which cause hospital infections, become stronger and less responsive to antibiotic treatment over time. Because of these characteristics, they require advanced and expensive tests for diagnosis, expensive antibiotics for treatment and thus lengthened hospital stays.

In order to control and prevent hospital infections, surveillance methods, which comprise systematic data collection, analysis, interpretation and reporting, are employed. According to the results of these surveillance methods, infection control programs are created, applied and monitored. In some medical centers, hospital information systems are used and through the data provided by these systems, hospital infections can be traced online with more controlled and extensive surveillance methods. Some of these systems adopt data mining applications to detect outbreaks, which cannot be determined easily by infection control teams.

Both descriptive and predictive data mining techniques can be exploited by hospital infections. Using lab data including test results for each department, descriptive techniques can be used to build models that produce early warning signals for outbreaks or new bacteria. On the other hand, predictive techniques can be used to determine risky patients which allow early intervention opportunity before the patient is infected or before applying unnecessary tests and treatments to noninfected patients.

In this study, predictive data mining techniques CHAID, CART and entropy decision trees, neural networks and logistic regression are applied on the newborn hospital infections data set collected by the Department of Clinical Microbiology and Infectious Diseases in Eskişehir Osmangazi University, Faculty of Medicine. Knowledge Discovery in Databases (KDD) methodology is followed throughout the study. Models are built with holdout sampling and cross validation methods and compared regarding accuracy and sensitivity measures of the test data set. Besides these measures, specificity, area under ROC curve, gini coefficient and average squared error measures are taken into consideration. In addition, bagging and boosting methods are applied on neural network and decision trees in order to increase the accuracy of these models.

The organization of the study is as follows: In Chapter 2, the literature on hospital infections in newborn intensive care units and data mining applications in the field are overviewed. In Chapter 3, data mining methodology, data mining techniques used in the study and important features of medical data mining are provided in detail. In Chapter 4, problem definition is presented, data preprocessing steps are explained. In Chapter 5, model results with applied accuracy estimation and increasing methods are analyzed. Finally in Chapter 6, the conclusions drawn from the study and possible further research directions are expressed.

CHAPTER 2

LITERATURE REVIEW

This chapter introduces literature review about hospital infections on newborn infants and data mining applications for hospital infection control. Appendix A comprises medical terms mentioned in this study.

Hospital Infections

Hospital infections or nosocomial infections are infections that originate or occur in a hospital or in a health care service unit. They are not in incubation period when the patient is admitted to the hospital and first appear after admission in 48 hours or after discharge within 10 days. Despite the improvements in hospital services, they can be seen in both developing and developed countries with increasing morbidity. They may cause functional disorder, emotional stress, decrease in quality of life, and death. In addition they increase hospital costs by lengthened hospital stay, antibiotics usage, isolation needs and other additional treatment methods (Ertek, 2008).

In a study conducted at a university hospital in Turkey, researchers found higher device-associated infection rates in intensive care units (ICU) than those reported by National Nosocomial Infections Surveillance (NNIS) System in United States. It is also reported that an infection control program is required to reduce the rate of device-associated nosocomial infections (Usluer et al., 2006).

As reported by Perk (2008), the risk of nosocomial infections on newborn has been increased in recent years, because of various invasive methods that are used to increase the living rate of newborn with very low birth weight (VLBW). Before birth, a newborn does not have a flora that protects him from colonization with harmful microorganisms. Moreover, prematures are immunologically immature and very open to any infection. Some of the risk factors can be listed as follows: premature birth, low gestational age, VLBW, invasive methods (mechanical ventilation, catheters), antibiotics and steroids usage, parenteral feeding, lipid usage, and the population in neonatal intensive care units (NICU). NICUs differ in the percentage and type of nosocomial infections regarding the risk level of neonates. In NICUs of neonates with high risk level, blood related infections are seen and the percentage of nosocomial infections is higher than those with low risk level. With low risk premature neonates, skin and mucosa colonizations are widely met.

According to a study of Borghesi and Stronati (2008); immunological immaturity, frequent use of invasive procedures and prolonged hospitalization account for the high incidence of infection in a population subgroup of 21% of all VLBW infants and up to 43% of neonates with birth weights of 401 - 750 g proven sepsis (sepsis with positive blood culture) was developed. They have also introduced strategies for the prevention of nosocomial infections such as hand hygiene practices, prevention of central venous catheter (CVC)-related bloodstream infections (CRBSI), and judicious use of antimicrobials for therapy and chemoprophylaxis, enhancement of host defenses, skin care and early enteral feeding with human milk.

In the study of Sameshima and Ikenoue (2006) examining newborn infants exposed to intrauterine infection, it is introduced that premature infants are more susceptible to intrauterine infection to cause death or cerebral palsy than mature infants. The critical gestational age for death (<28 weeks) is younger than that for cerebral palsy (<34 weeks) with these infants. It has been found that both infection and acidosis were required in mature infants to cause cerebral palsy whereas infection alone can cause brain damage in less mature infants.

A study of van Rossem et al. (2007) shows that colonization of neonates with Enterobacter spp. (species) does not usually lead to clinical infection. Colonization of neonates treated in a NICU of a tertiary care hospital with Enterobacter spp. was associated with prolonged antibiotic use and low Apgar scores, reflecting the severity of underlying illness.

The results of Matussek, Taipalensuu, Einemo, Tiefenthal and Löfgren's (2007) study confirms a high level of transmission of Staphylococcus aureus from staff members to infants (13 cases of 44) compared with parents (11 cases of 44) and spa typing epidemiological tool is suggested to improve hospital hygiene control programs.

Septicemia is the most common neonatal infection in the NICU (45-55%), followed by respiratory infections (16-30%), and urinary tract infections (8-18%) (Clark, Powers, White, Bloom, Sanchez, & Benjamin, 2004)

It is claimed by Pittet (2005) that annually 5 to 15 % of in-patients are being diagnosed with hospital infections in US, which inferred to ca. 2,000,000 diagnosis. 44,000 to 98,000 of those result in death. Infections also add an estimated $17 to $29 billion to the US's hospital costs annually. 25 to 50 % of hospital infections are observed in ICUs in US.

Görenek (2002) reported that hospital infections have been seen on 3.1 to

14.1% of in-patients in Turkey. Hospital infections bring $1,500 cost per patient at an average of two weeks lengthened hospital stay. For pediatric patients this cost can reach up to $10,000. The risk of infection is 8 to 10 times higher by ICUs in comparison with others. Effective surveillance and appropriate actions can reduce this risk only by 20-30%.

These statistics represent the importance of hospital infections and witness high mortality and morbidity, lengthened hospital stays and increased hospital costs. Thus preventive programs have to be developed for this public health care problem.

Very effective programs were shown to reduce infection by 32% in the study of the Efficacy of Nosocomial Infection Control (SENIC (1970-1975)). Among key results from the SENIC project was the information that between 35% and 50% of all nosocomial infections were associated with a few patient care practices: use and care of urinary catheters; use and care of vascular access lines; therapy and support of pulmonary functions; experience with surgical procedures; and appropriate hand hygiene and use of isolation precautions (Pittet, 2005). According to SENIC results, the most important basic elements of handling nosocomial infections are: nosocomial infections surveillance and its control activities, a control team of nosocomial infections experts with enough number of members according to the size of the health care institution, and finally sharing the results of surveillance with related parties (Erol, 2008).

In the following section, how data mining is adopted in such surveillance systems and the success in those systems are reviewed.

Data Mining Applications in Infection Detection

The Infection Control Department at St. Luke's Episcopal Hospital is committed to provide high quality health care to the community they serve. In order to fulfill their commitment, they implement technology to collect and store data and applied data mining techniques (Dao, 2006).

Pittet (2005) referred to data mining derived epidemiology as one of the major challenges for future: Fully computerized patient records bring new opportunities for the development of "at risk" patient profiles, thus prompting earlier intervention strategies. It may also allow for data mining to help sort patient characteristics associated with higher or specific risks for health care-associated complications and, in particular, infection.

According to the January 2004 automated data mining surveillance system (DMSS) report in Saint Francis Hospital, Memphis, Tennessee, a mini-cluster of four Escherichia coli (EC) urinary isolates related to patients originating from the orthopedic unit was found. The finding was investigated and urinary catheter selection in emergency room was found to be the reason. Infection control specialists changed the program and DMSS resulted in an improvement chart with 3 consecutive months of zero EC urinary isolates. (Breaux, Baker, Wilburn, Monteith, & Umstead, 2005).

Kreuze (2001) cites DMSS success in a 600-bed tertiary-care medical center in Alabama, US. The system has detected 41 suspected outbreaks, subsequent inspection of patients' charts revealed that 97% actually had hospital infections. During that same period, the medical center's infection control staff has flagged only

9 potential outbreaks of which three of them were turned out to be true.

Effective data mining permits microbial pattern recognition and detects the presence of microbial clusters that warrants early investigation to enhance patient safety and prevent costly outbreaks (Dao, Zabaneh, Holmes, Disrude, Price, & Gentry, 2008).

In the study of Meek and Tinney (2006), it is reported that data from hospital databases including patient admissions and results of lab tests are analyzed for infection trends. Before nosocomial infections spread, the data mining service in the hospital looks across the entire facility for early indications.

As high risks and costs are associated to nosocomial infections, in order to monitor them in various areas of the hospital and to identify and report the critical situations for patients, Lamma, Manservigi, Mello, Storari, and Riguzzi (2000) developed a descriptive system in which clustering algorithms are used. They compute the frequency of infections and expect them to highlight possible hygienic problems and to be used for early diagnosis and therapy over time. The system also generates alarms regarding newly identified bacteria: when an unexpectedly resistant bacterium is found, when a contagion among patients of a unit is detected, or when the therapy is found to be ineffective.

CHAPTER 3

METHODOLOGY

In this chapter, knowledge discovery process, data mining techniques used in the study, model accuracy, techniques for evaluating and increasing the accuracy and finally important features of medical data mining that need attention are discussed.

Knowledge Discovery in Databases

The data that build up the information can be in a complex structure and also be incomplete, inconsistent, incomparable, extreme or even unnecessary. In order to obtain useful knowledge from these data for decision support; knowledge discovery in databases process (KDD process) is used.

The process of KDD has been defined by Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996) as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (pp.6). According to this definition, features of a pattern can be expressed as follows. A pattern is “valid”, if the degree of certainty is acceptable, when applied on a new data set. A pattern is “novel”, when compared to previous values or knowledge. A pattern is “potentially useful”, as measured by some utility function, as an increase in profits. Finally, a pattern is needed to be “ultimately understandable”, which can be measured by the simplicity of the pattern, in order to facilitate a better understanding of the underlying data.

KDD process covers mainly goal identification, data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation steps. Out of these steps, data mining is concerned with extracting a pattern with relevant features from large amounts of data under some computational limitations. Though it is the core step in KDD process, it can be also used as a synonym for KDD (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). Figure 1 represents the basic flow of KDD process with interactions between the steps.

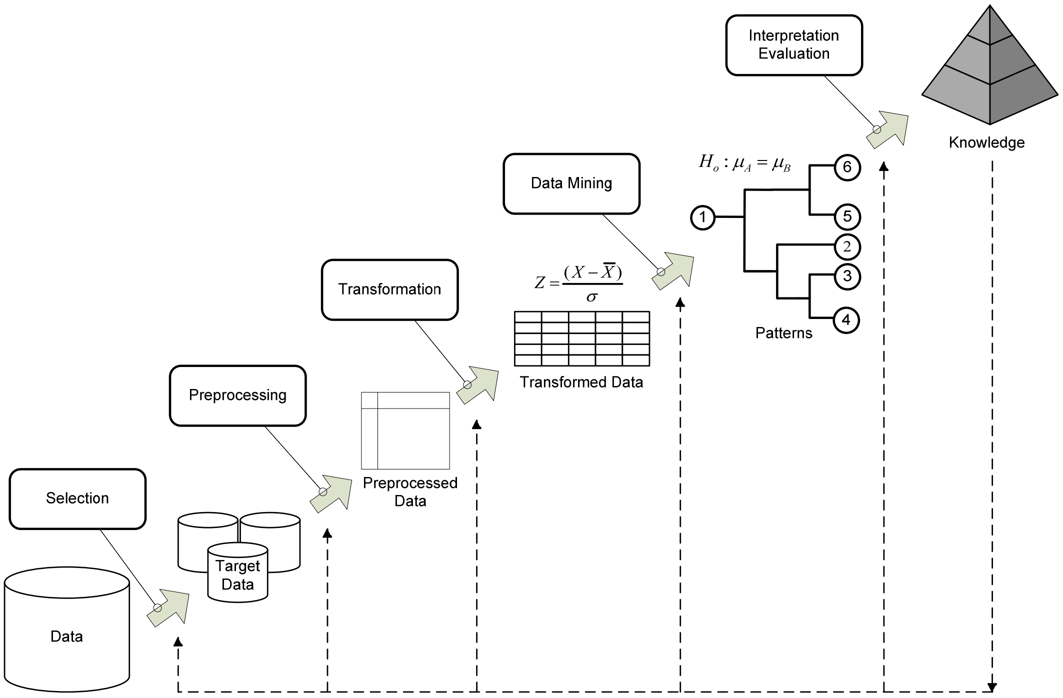

Figure 1. Basic flow of KDD process

KDD process is an interactive and iterative process, meaning, many decisions about parameters, measures and thresholds are met by user and loops are allowed between any steps in order to constitute a meaningful knowledge at the end of the process.

The process is first defined by Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996), and comprises the following steps:

1) Developing an understanding of the application domain: Gathering information about application domain, relevant characteristics and other prior knowledge. Understanding end-user goals.
2) Creating a data set: Selecting and integrating variables of the data set in order to use in the discovery process.
3) Data cleaning and preprocessing: The removal of noise or outliers if appropriate, and deciding the strategies for handling missing values regarding the characteristics or changes in the source.
4) Data reduction and transformation: Depending on the goal of the data mining task, from a list of variables, relevant ones are chosen and transformed in order to increase the relevancy.
5) Choosing appropriate data mining task: Regarding the goal of the KDD process, appropriate data mining task is selected.
6) Choosing the data mining algorithm: The specific method with its appropriate parameters is chosen to be used for searching patterns in line with the (expected features of result) consideration. For instance, if end-user needs to understand the pattern, not to have a precise one, then decision tree algorithms can be used instead of neural networks.
7) Data mining: Employing the algorithm and generating the patterns; if necessary, adjusting its parameters in order to obtain satisfied results.
8) Evaluation: Pattern evaluation and interpretation regarding predefined goals. Previous steps are considered to be repeated considering their

effects on data mining results and patterns are evaluated in terms of comprehensibility and usefulness.

9) Using the discovered knowledge: Deploying the data mining algorithm into performing systems to take actions, or report the results to end-users for them to compare the results with previously known. The effectiveness of the entire KDD process can be acquired at this step. A successful pattern will face several challenges in this step, such as a dynamic environment with changes in variables either because of data structure or data domain.

The first four steps are considered as goal definition and data preprocessing. The success of the process mostly depends on the quality and quantity of the data. Therefore, data preprocessing needs extra effort and covers an important portion of the process. The consequent three steps represent the preparation for data mining and applying data mining. Finally, the last two steps are the evaluation and interpretation of the results and monitoring the model after it is deployed into a performing system.

After Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996) described this first KDD process model, Anand and Buchner (1998) developed an eight step model. Other than these academic research models, industry specific models were also developed. The most known and used industry specific models are the "Cabena, Hadjinian, Stadler, Verhees, and Zanasi" (1998) model, developed with the support of IBM and the "CRISP-DM" (CRoss-Industry Standard Process for Data Mining) model, developed by a large consortium of European companies (CRISP-DM Web, n.d.).

Next section explains data mining, the core step of KDD process in more detail. For this, the classification of data mining techniques are defined and summarized. In following sections, a comprehensive overview of data mining techniques that were used in the study and methods for estimating and increasing classifier accuracy are introduced.

Data Mining

Data mining is a step in KDD process that refers to extracting or "mining" knowledge from large amounts of data (Han & Kamber, 2006). For specified data mining tasks, relevant data mining techniques are applied in order to find novel and useful patterns that might be unknown, if not searched with these techniques.

Data mining tasks are generally classified into two broad categories as descriptive and predictive. Descriptive data mining tasks are exploratory tasks and they search for the general properties like trends, correlations, clusters, and anomalies representing the underlying relationships in the data. On the other hand, predictive data mining tasks make predictions for a variable called dependent variable or target, based on other variables called independent or input variables.

Two types of predictive data mining tasks can be performed: classification and prediction. Classification is the process of finding a model that describes and distinguishes data classes, for the purpose of predicting the discrete target variable whose class label is unknown. Decision trees, neural networks, and logistic regression are examples for classification techniques. Whilst classification predicts discrete (categorical) labels, prediction models continuous-valued functions: missing or unavailable numerical data values (Han & Kamber, 2006). Regression analysis is an example for continuous target models.

Association analysis, cluster analysis, outlier analysis and evolution analysis are groups of different descriptive data mining tasks.

Patterns that occur frequently and describe strongly associated features in data are called frequent patterns. Association analysis discovers the patterns of associations and correlations represented in the form of implication rules in sequences or frequent item sets.

Unlike classification and prediction, clustering analyzes the data without a known dependent variable and tries to find groups of closely related observations based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. In other words, clustering tries to form clusters of objects that are homogenous inside and heterogeneous between each other.

Outlier analysis or anomaly detection identifies observations that have significantly different characteristics from the general behavior of data. These are called outliers or anomalies and they are eliminated from most data mining methods as noise or exceptions. However, the goal of this analysis is to discover the real outliers and avoid assigning normal observations as outliers. Outlier analysis is used in fraud detection, network intrusions and unusual patterns of disease.

The above mentioned techniques are mostly used on simple and structured data sets, such as data extracted from relational databases and data warehouses. Complex forms of data like spatial and temporal data, hypertext and multimedia data and semi-structured or unstructured data needs to be mined with more advanced data mining techniques. Evolution analysis comprises the analysis of such data and models regularities or trends for observations whose behavior change over time: Time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis (Han & Kamber, 2006).

Classification Techniques Used in the Study

The goal of the study is to classify nosocomial infection on new born infants based on its relevant characteristics. This section introduces the classification techniques *decision trees*, *neural networks*, and *logistic regression* in more detail that are used in the study.

Decision Trees

Decision trees are one of the most popular classification and prediction methods used in data mining. As mathematical inference techniques are hard to be understood by users, decision trees are developed by machine learning researchers with the purpose of providing ease of human use and interpretation.

Han and Kamber (2006) defined decision tree as a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

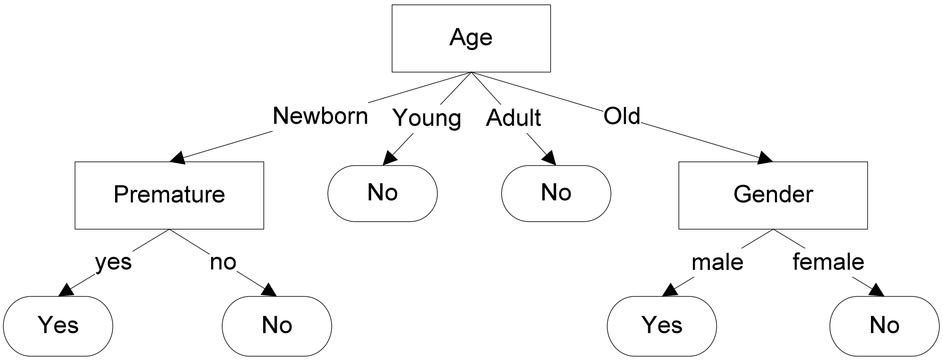

Figure 2. Decision tree representing the presence of hospital infections.

Figure 2 describes a decision tree structure that represents the presence of hospital infections on patients. Each internal node represents a test on an attribute (premature or not, male or female) whereas each leaf represents a class (infected or not).

Classification rules can easily be derived from decision trees by starting from the root node to a leaf node that suggests the class prediction. In the above example, the classification rule "If the patient is a newborn and premature, then the patient is likely to be diagnosed with hospital infections." can be derived from the decision tree structure.

Decision trees are mostly preferable because they do not require any domain knowledge or parameter setting that complicates the construction; besides they provide easy implementation and integration to databases. They can also handle several types of variables like nominal, interval, and ordinal. In addition to the main goal of prediction, decision trees can be used for interaction detection, stratification, missing value imputation, and model interpretation. They are also used for variable selection from a wide range of variables before applying other data mining methods like regression or neural networks.

The fundamental principle underlying tree creation is that of simplicity: Decisions that lead to simple, compact tree with few nodes are preferred (Duda, Hart,

& Stork, 2001). The tree complexity is also explicitly controlled by the stopping criteria used and the pruning method employed (Maimon & Rokach, 2005).

To be able to construct such an accurate decision tree, best splitting variable is searched, which will make the descendent node as pure as possible. For formulization, defining impurity is more convenient than purity of a node. Although there are several splitting measures having the same behavior of defining impurity, three of them will be discussed in this study: information gain, gain ratio and Gini index.

Information gain is the most popular measure originated from information theory. Input variable with the highest information gain is selected in order to decrease the information need of descendent nodes and represents their impurity. Let I(S) be the impurity of a node S and n(S) be the number of observations in node S. The information impurity (or entropy impurity) will be defined as follows:

$$I(S) = - \sum_{i=1}^{m} \mathrm{p_i} \log_2(\mathrm{p_i})$$

where p_i is the probability of class *i* in node S and is estimated by $n(S_i)/n(S)$. I(S) gives the average information value needed to classify the observations in node S, which is based on the proportions of observations of each class. The expected information to find an exact classification after further partitioning of node S, is measured as follows:

$$E(A) = \sum_{j=1}^{v} \frac{n(S_j)}{n(S)} \times I(S_j)$$

where *j* is the class label of new feature A until *v*, and S_j is the subsequent node of S. The smaller the expected information, the greater the purity of partitions.

The difference between the original information impurity and expected information is the information gain:

$$Gain(A) = I(S)-E(A)$$

Knowing the value of feature A will reduce the information requirement and the attribute with the highest information gain will be chosen for splitting.

Information gain has a drawback that it prefers to select features with large number of values. Because a feature with large number of values will provide more partitions and will reduce the impurity more than another one with small number of values will. For instance, a unique identifier will provide expected information of 0 and therefore the information gain will be maximal (Han & Kamber, 2006). To overcome this bias, gain ratio is introduced, which applies a kind of normalization to information gain using split information. The split information is defined as follows:

$$SplitI(A) = \sum_{j=1}^{v} \frac{n(S_j)}{n(S)} \times \log_2 \left(\frac{n(S_j)}{n(S)} \right)$$

For each outcome, the number of observations having that outcome with respect to the total number of observations in S is considered. On the other hand, information gain measures the information with respect to classification that is obtained based on the same partitioning. The maximum value of gain ratio is selected for splitting which can be calculated with the following equation:

$$GainRatio(A) = \frac{Gain(A)}{SplitI(A)}$$

Another measure of impurity is the Gini index that considers a binary split for each feature. For discrete features the subset, and for continuous features the split point that gives the minimum gini index is selected for splitting. Gini index is expressed as follows:

$$Gini(S) = 1 - \sum_{i=1}^{m} p_i^2$$

Although these splitting measures have some bias, they are used successfully in practice. However, growing a tree until the lowest impurity is met at leaf nodes will cause overfitting. Thus it is not expected for fully grown trees to generalize the noisy problems well. In return, the error will not be low enough and performance of the tree will be week, if splitting is stopped early. To overcome these problems, cross-validation can be used, which will be explained in further sections. Other than cross-validation, a stopping criterion (pre-pruning approach) can be used by defining a threshold value in advance. This approach avoids generating too complex subtrees, nevertheless, it is often difficult to set the threshold, because there is rarely a simple relationship between the preset threshold and the ultimate performance. Too high thresholds will result in underfitted models, while low values may not be sufficient to overcome the overfitting problem.

In pre-pruning approach, there is the risk to cut off the possibility of beneficial splits in subsequent nodes by meeting the stopping threshold too early. In post-pruning, pruning can be done by merging leaf nodes whose elimination provides a satisfactory increase in impurity to an antecedent node. Although this is the most common approach, in cost-complexity pruning, a complex subtree can be replaced with a leaf directly. Post- pruning approach is more likely to be used with small data sets because of computational expense, when building the decision tree.

Decision tree algorithms automatically derive decision trees from training data sets. Most decision tree algorithms adopt top-down recursive approach, which is also known as divide-and-conquer during the construction of a decision tree.

Algorithms start with partitioning the training data into subsets with the most appropriate splitting variable, according to some splitting measure. Until sufficient splitting measure is not fulfilled or one of stopping criteria is met, internal nodes continue to partition the training data.

ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), CART (Breiman, Friedman, Olshen, & Stone, 1984) and CHAID (Kass, 1980) algorithms are the examples of top-down approach performing decision tree inducers. ID3 and C4.5 algorithms were developed in the field of machine learning, whereas CART and CHAID were developed in the field of applied statistics.

The ID3 algorithm is a simple decision tree algorithm, which adopts information gain as a splitting criterion and does not perform any pruning method. The decision tree stops growing when information gain is not greater than zero or all the observations in a node belong to the same class. This algorithm also does not work with interval and missing values. As a successor of ID3 algorithm, C4.5 algorithm uses gain ratio as a splitting criterion and error based pruning. It can handle both numeric and missing values.

The CART (Classification and Regression Trees) algorithm generates binary trees using Gini and twoing splitting criteria and cost-complexity pruning. For interval targets CART algorithm constructs regression trees and reduction in squared error is used as splitting criteria.

The CHAID (Chi-Squared Automatic Interaction Detection) algorithm uses statistical Pearson chi-square for nominal targets, likelihood-ratio for ordinal targets and F test for continuous targets, nevertheless does not perform any pruning method. Missing values are handled as a single category. CHAID evaluates all values of an input variable and derives least significantly different (the greatest p value) pair of values concerning target variable. If the p value is greater than a previously specified merge threshold, than CHAID merges these values into a single category. After adjusted p values are computed, input variable with the smallest adjusted p value is selected and compared with a previously specified splitting level. If p value is smaller or equal to the splitting level, splitting continues with the related input variable, otherwise no split is performed. CHAID also stops when the stopping rules such as maximum tree depth, minimum number of observations in an internal node or minimum number of observations in a leaf node are met.

Neural Networks

Neural nets or networks are relatively simple mathematical models that simulate the biological nervous system. The common characteristics of neural networks and biological neurons are parallel processing of information and learning and generalizing from experience. Neural networks are used for many data mining tasks, such as pattern classification, time series analysis, prediction and clustering.

Neural networks can be used when there is a poor relationship between inputs and outputs and they are also successful on incomplete and noisy data and classifying patterns on which they have not been trained (Han & Kamber, 2006). As a classifier type, neural networks are mostly nonparametric unlike the normal linear classifiers. No assumption is made about the underlying population distribution.

Han and Kamber (2006) described neural networks as “a set of connected input/output units where each connection has a weight associated with it”. Network learns by adjusting the weights to predict the correct class label of the sample cases.

The simplest neural network device is single layer perceptron with only one input and one output layer of neurons. It is the equivalent of a linear discriminant that is simply a weighted scoring function and successful only when the two classes are linearly separable (Weiss & Kulikowski, 1991). As an extension to single layer perceptron, multilayer perceptron is introduced, which overcomes the restriction of linearly not separable problems and consists of one input, one or multiple hidden layers, and one output layer. Figure 3 (Tan, Steinbach, & Kumar, 2006) represents a multilayer feed-forward network with one hidden layer.

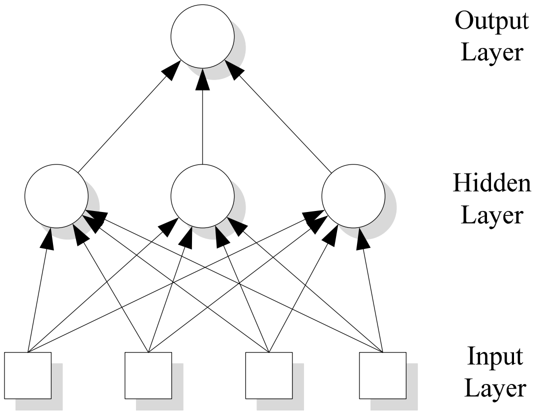

Figure 3. Multilayer feed-forward perceptron with one hidden layer

Feed-forward networks consist of nodes that are connected only to the nodes in the next layer. None of the weights returns back to an input or output node of a previous layer. On the other hand, in recurrent networks nodes in the same layer can be connected to each other.

When designing a network topology, number of nodes in input layer is determined by the type of input variables. Each numeric variable is normalized to speed up learning and assigned to an input node. Besides each character variable value is encoded as binary variables and like other binary variables each of these are assigned to one input node. Only one node in output layer is sufficient if the target is binary. If the target variable has *n*-classes, then the number of output nodes is *n*. To find correct number of hidden layers is a trial-and-error process and may affect the accuracy of the network. The initial values of weights and biases, which are chosen randomly also affects the accuracy. Therefore, small random numbers will be appropriate.

There are several types of neural network algorithms. Here the most popular one, backpropagation, is referred, which performs on multilayer feed-forward networks. Figure 4 illustrates the forward processing of information in a hidden or output node.

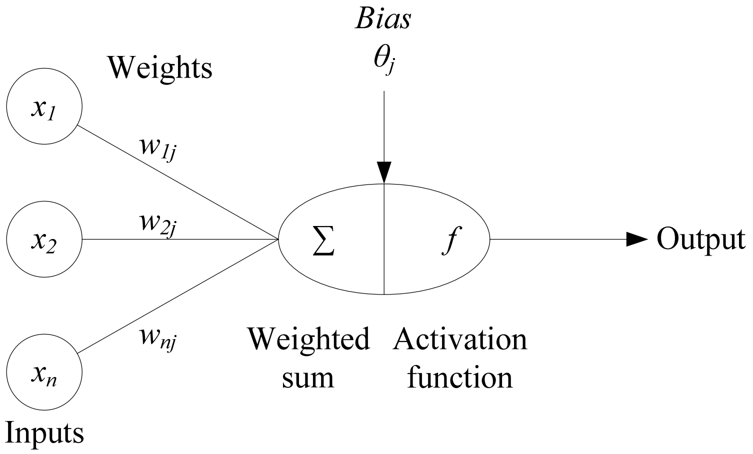

Figure 4. Process in a single hidden (output) layer node

Backpropagation algorithm processes each training observation sequentially and works iteratively. It starts with propagating the inputs forward, feeding the training observation to the input layer and assigning initial weights and bias for the first iteration. For an input layer node *j*, the output O_j is equal to the actual input value I_j ($O_j = I_j$). For hidden or output layers, the input and output value of each node are computed with the weights obtained from the previous iteration. The input I_j is

$$I_j = \sum_i \mathrm{w}_{\mathrm{ij}} \, O_{\mathrm{i}} + \theta_j$$

where w_{ij} represents the weight of connection from node *i* in the previous layer to node *j*, O_i represents the output of node *i* from previous layer, and θ represents the bias. Bias is a threshold that varies the activity in the node. In order to compute the output values that are nonlinear, a logistic activation (transfer) function is applied to input I_j .

$$O_j = \frac{1}{1+e^{-I_j}}$$

where O_j is the output of the logistic activation function applied on input I_j. As the goal is to minimize the mean squared error between predicted and actual target value, the weights are modified in backward direction, starting from output layer through the hidden layers and input layer. First, the error ($Error_j$) for each output layer node *j* is computed

$$Error_j = O_j(1-O_j)(T_j - O_j)$$

where O_j is the predicted and T_j is the actual value of output. Second, the error in each hidden layer output is computed. Therefore, from the last to the first hidden layer, the weighted sum of the errors of the next higher layer node are taken into consideration.

$$Error_j = O_j(1-O_j)\sum_k Error_j w_{jk}$$

where w_{jk} is the weight of the link between node j and the node k in previous layer and $Error_k$ is the error of node k.

Finally, the weights and the biases are adjusted. While backpropagation learns by a method of gradient descent, there is a risk to meet local minimum instead of global minimum when computing the mean square error between prediction and actual. Therefore, a constant variable *l*, namely *learning rate* is used to compute the change. Weights are adjusted as follows

$$\Delta w_{ij} = (l)\sum_{k} Error_j O_i$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

where Δw_{ij} is the change in weight w_{ij} and biases are adjusted as follows

$$\Delta \theta_j = (l)\sum_{k} Error_j$$

$$\theta_j = \theta_j + \Delta \theta_j$$

where $\Delta\theta_j$ is the change in bias θ_i.

One iteration through all observations of a training set is called an epoch. In theory, adjustments are made after all observations of the training set are processed and this approach is called *epoch updating*. However in practice, adjustments after training each observation give more accurate results.

Backpropagation stops training when the changes in weights are smaller than a predefined threshold, misclassification rate is below a predefined threshold or maximum number of iterations has reached.

Neural networks are poorly interpretable, because the importance of independent variables is difficult to be measured as independent variables are associated with both hidden and output layer weight estimates. These estimated weights don’t explain the degree or rate of change in the relationship between the dependent and independent variables in the model. Hence, several methods were developed to extract rules from trained networks. These can be grouped into following categories (Chakraborty and Chakraborty, 1999):

- Algorithms that extract rules by analyzing the effect of an individual input on an individual output as a whole.
- Algorithms that examine the effect of the inputs on activation of the hidden units and extract rules by clustering inputs.
- Algorithms that try to interpret the connection weights to find out decision rules.
- Algorithms that extract rules by analyzing the function learned by the neural net and the effects of the input on learned function.

Logistic Regression

Logistic regression is a predictive modeling technique, a special form of regression, which is developed for binary categorical targets. It finds out the relationship between two or more independent variables and a single binary dependent variable.

Logistic regression differs from multiple regression by the type of dependent variable, estimation methods, and assumptions about the underlying distribution (Hair, Anderson, Tatham, & Black, 1998). It violates two assumptions of linear regression: Instead of a normal distribution, the error term of discrete variables follows the binomial distribution and the variance of a binary variable is not constant.

On the other hand, logistic regression differs from multiple discriminant analysis, another multivariate technique, by being less affected when the basic assumptions, particularly normality of the variables are not met (Hair, Anderson, Tatham, & Black, 1998).

Logistic regression predicts the probability of an event occurring, however to hold the predicted value in a range of zero and one, it uses an S-shaped logistic curve that symbolizes a relationship between independent and dependent variables.

Figure 5 illustrates a logistic curve drawn for a single independent variable *x*. At very low values of the independent variable, probability approaches to zero. For a certain range of intermediate values, probability increases rapidly and at very high values, it approaches to one, but never exceeds these boudaries.

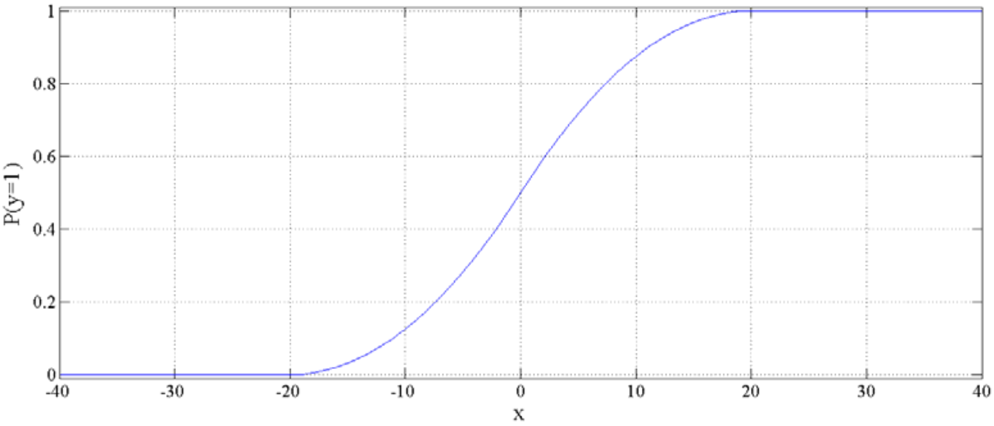

Figure 5. Logistic curve

Given several independent variables, logistic function is defined as the following equation,

$$p_i = \frac{1}{1+e^{-(\alpha+\beta_1\chi_1+\beta_2\chi_2+...+\beta_j\chi_j+...+\beta_n\chi_n)}}$$

where α is the intercept, β_j is the coefficient of independent variable x_j and p_i is the probability of the binary dependent variable *y* being 1. $P(y = 1) = p_i$, $P(y = 0) = 1 - p_i$

In order to obtain the unknown parameters, the intercept (α) and the coefficients (β_j), logistic regression uses maximum likelihood method because of its nonlinear structure. Therefore, likelihood value is used to calculate the measure of overall model fit. On the other side, multiple regression mostly uses least squares method to estimate the coefficients and sum of squares for the measure of overall model fit.

To transform the values of binary dependent variable into logistic curve that represents the probability of an event, logit transformation is used.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_j\chi_j + \ldots + \beta_n\chi_n$$

logit transformation is the log of odds ratio, which is an important measure for logistic regression that compares the probability of an event with the probability of an event not happening, and is calculated as follows:

$$\text{Odds ratio (OR)} = \frac{p}{1-p}$$

The estimated coefficient β_j in logit function represent the changes in log odds ratio for a unit change in χ_j. Positive coefficient values increase; negative coefficients decrease the predicted probability. For instance, if β_j is positive, its transformation will be greater than 1, the odds ratio will increase and therefore, the model will predict y = 1 better.

In order to test the significance of estimated coefficients, Wald statistic is mostly used (Hair, Anderson, Tatham, & Black, 1998). It can be interpreted like F or t values used for significance testing of regression coefficients. The Wald statistic is not used when the coefficient is extremely large, especially when there is a binary independent variable with a mean closer to 0 or 1. In such situations, likelihood ratio or score statistic, a measure of association in logistic regression, should be used.

Logistic regression maximizes the likelihood of the dependent variable being 1 (y = 1), whereas multiple regression minimizes the squared deviations. Likelihood measure is an overall measure of goodness of fit and used as -2 times the log of the likelihood value. It is referred to as -2 log likelihood or -2LL. The smaller values of -2LL indicate good models. The change in -2LL with chi-square test and significance also provide a measure of improvement in model when a new independent variable is added. In order to measure the change, a Pseudo R-Square is calculated by using the information from the -2LL for the full model and the intercept only with the following equation:

$$Pseudo\ R^2 = \frac{(\text{-2LL}_{\text{null}}) - (\text{-2LL}_{\text{model}})}{(\text{-2LL}_{\text{null}})}$$

where $\text{-2LL}_{\text{null}}$ is -2LL of only the intercept model and $\text{-2LL}_{\text{model}}$ is -2LL of the final model (Hair, Anderson, Tatham, & Black, 1998). This pseudo R-Square is known as McFadden's R-Square and it can be as low as zero but can never be one.

Another measure of assessment is the Hosmer & Lemeshow measure of overall model fit. This measures the correspondence of the actual and the predicted values of dependent variable. For this measure, the greater is the significance value, the more similar are the observed and predicted classifications.

Classifier Accuracy

This section summarizes classifier accuracy and methods for estimating and increasing the accuracy. Classification models are built on a training set that consists of records whose class labels are known, and if needed, a validation set can be used for model fine-tuning (e.g. for pruning a decision tree). Most classification algorithms seek for models, which perform successfully on previously unseen test data set. The success is measured either by classifier's accuracy or error rate. Accuracy reflects the overall correctness of the classifier, and calculated as follows.

$$Accuracy = \frac{number\ of\ correct\ predictions}{number\ of\ observations}$$

The error rate is the ratio of the number of errors to the number of observations, where the error is simply misclassification, the number of cases that the classifier classified incorrectly.

$$Error\ Rate = \frac{number\ of\ errors}{number\ of\ observations}$$

Error rate is also calculated as *1- Accuracy.*

As the classification models are built on a training set, error represents only the misclassification on the training set and is called the training error. The expected error on previously unseen test data set is called generalization error. A good model is expected to have both low training and low generalization error. If a model with low training error has a higher generalization error than a model with higher training error, then this is called overfitting (Tan, Steinbach, & Kumar, 2006). When a model is overfitted, after a certain point, generalization error starts to increase, while training error continues to decrease. If a model has not learned the pattern well, and performs poorly, then it has high training and generalization error rates and this is called underfitting (Tan, Steinbach, & Kumar, 2006). Both overfitting and underfitting are originated from model complexity.

In order to build a model that has the appropriate level of complexity to avoid overfitting, lowest generalization error is sought. Hence, several methods for estimating the generalization error are used during training. After the model is built, it can be tested on a data set, which is unknown to training model before.

In some fields like medicine, the distinctions among different types of errors are important. For such cases, confusion matrix that lists the actual against predicted classification can be used. Figure 6 represents a confusion matrix, where a binary target is modeled.

		Actual	
		1	0
Predicted	1	True Positive (TP)	False Positive (FP)
	0	False Negative (FN)	True Negative (TN)

Figure 6. Confusion matrix for binary target

In medicine, false negatives and false positives are not treated equally, especially in life-threatening illnesses. Diagnosing a healthy person with cancer is expected to be determined later with further tests however, treating a cancer patient as healthy may cause death. Besides, accuracy rate of 90% does not provide much information, where no information about the true positive rate is given. Therefore, measures like sensitivity and specificity are widely used, which are derived from confusion matrix (Tan, Steinbach, & Kumar, 2006). Sensitivity is the proportion of positive observations, whereas specificity is the proportion of negative observations that are correctly classified. The formulas of sensitivity and specificity regarding confusion matrix are as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

A graphical approach that shows the trade-off between the sensitivity (also known as true positive rate) and 1-specificity (false positive rate) for a given model is the

Receiver Operating Characteristic (ROC) curve (Han & Kamber, 2006). ROC curve also allows for comparing the relative performance among several classifiers. Figure 7 shows the ROC curves of classifiers M_1 and M_2.

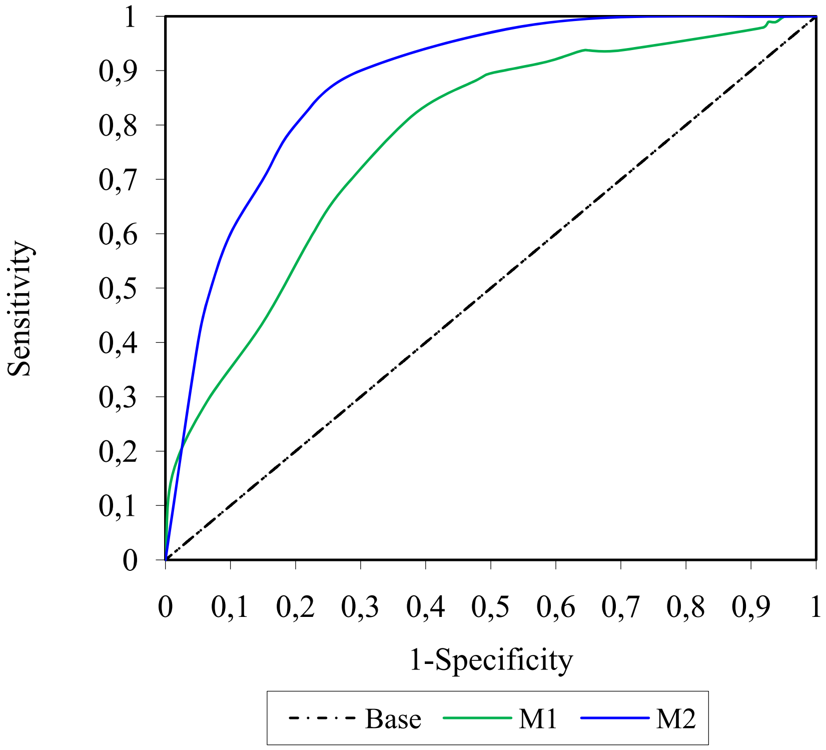

Figure 7. ROC curves for classifiers M_1 and M_2

Figure also represents a diagonal base line, where for every true positive; it is equally likely to find a false positive. This could be a model that makes random guesses. Hence, the closer the ROC curve of a model to this base line, the less accurate the model and it is suggested to operate always above this line, where true positive rate is always greater than false positive rate. An ideal model would be at the point where true positive rate is 1 and false positive rate is 0. So, a good model should be as close as possible to the upper left corner of the diagram. Area under ROC curve (AUC) is also used to assess the accuracy of a model and is equal to 1 if the model is perfect, equal to 0.5 if the model makes random guesses.

Costs and benefits can also be incorporated by computing the average cost and benefit per decision using confusion matrix. A misclassification cost is simply a number that is assigned as a penalty for making a mistake. By assigning different costs to different types of error, they can be outweighed over another. For instance, in cancer diagnosis false negatives can be assigned a cost twice the false positives. The same logic can also be applied for benefits, where benefit numbers reward the true predictions.

Accuracy Estimation Methods

Training error rate poorly estimates the future performance and tends to be biased optimistically. On the other hand, partitioning the data into train and test samples, and measuring the performance of the classifier on test sample provides unbiased estimates of generalization error. Accuracy and error rate of test samples can also be used to compare different classifiers. In order to make an overview of the methods to assess the accuracy *holdout, random subsampling, cross-validation,* and *bootstrap* methods are introduced in this section.

Holdout method is a basic method for evaluating the classifier’s accuracy. Data set is partitioned into two subsets, training and test respectively. Typical ratio for partitioning is 2/3 for training and 1/3 for test set (Weiss & Kulikowski, 1991). The classifier is trained on training set and its accuracy is estimated with test set. Holdout method is not suitable for small samples, considering the insufficient sizes of subsets for both training and test.

Random subsampling method lets the classifier run on more than one different train and test samples and provides the average of accuracies as the overall accuracy estimate. It can produce better results than a single train and test partition.

Cross-validation method basically uses each record the same number of times for training and exactly once for testing. It has three different variations and all of them are iterative (Tan, Steinbach, & Kumar, 2006). In k-fold cross-validation, data is partitioned into k equal-sized partitions. In each iteration a different partition is selected as test sample and the classification algorithm derives the pattern from the rest. The average of the accuracies that are derived from each iteration provides the overall accuracy. A special case of k-fold cross-validation, where k is equal to the number of sample cases, is leave-one-out cross validation. Only one sample case is used for testing in each iteration. Although it is computationally expensive, it is almost an unbiased estimator of the generalization error. In stratified cross-validation, the partitions are stratified so that the class distribution of target variable is approximately the same as the original sample. The great advantage of cross-validation method is that all cases in original sample are used for test, and almost all of them are used for training. While for small samples (ca. 30 cases) leave-one-out method is suggested, for moderate and large samples 10-fold cross-validation is suggested.

Unlike the methods mentioned before, bootstrap method samples the training set with replacement. Sampling with replacement means that the training cases are drawn from the data set and put back into the original data set after they are used. Thus duplicate records are allowed in the training data set. .632 bootstrap is a commonly used bootstrap estimator. When the original data set has N observations, it is shown that a bootstrap with N observations contains 63.2% of all the observations in original data set. The observations that are not included in bootstrap sample are used as the test sample and the sampling procedure is repeated k times. The overall accuracy is computed by combining the accuracies of each bootstrap sample ($Acc(M_i)_{test}$) with the accuracy computed from a training set that contains all the labeled examples in the original data set ($Acc(M_s)_{train}$) where M represents the classifier. The formula for overall accuracy is as follows:

$$Acc(M) = \frac{1}{k}\sum_{i=1}^{k}(0.632 \times Acc(M_i)_{test} + 0.368 \times Acc(M_s)_{train})$$

Although bootstrap method is not always superior to leave-one-out cross-validation method on small data sets, it is preferred while leave-one-out method has a high variance.

Accuracy Improving Methods

In addition to accuracy evaluation methods, this section introduces the methods that increase the accuracy of a classifier. Accuracy of decision tree classifiers can be increased with pruning. Other than decision trees, in general, ensemble methods like bagging and boosting help to increase the accuracy by combining a series models to create an improved composite model.

Bagging is also known as bootstrap aggregation and uses sampling with replacement method to create training and test samples, which makes it less susceptible to overfitting when applied to noisy data (Han & Kamber, 2006). One or several classification algorithms are applied on each training set and return their class predictions. Each prediction is taken as a vote, and majority vote is assigned as the final class by the bagged classifier. Bagging can also be applied to continuous valued targets by taking the average value of each prediction for a given test observation. Figure 8 (Tan, Steinbach, & Kumar, 2006) represents the basic procedure for bagging algorithm.

```
Bagging Algorithm
(1) Let $D_i$ be a bootstrap sample, $k$ be the number of bootstrap samples.
(2) for $i = 1$ to $k$ do
(3)     Create a bootstrap sample $D_i$.
(4)     Train a base classifier $M_i$ on $D_i$.
(5) end for
(6) $M^*(x) = \mathrm{argmax}_y \Sigma_i \delta(M_i(x) = y)$.
{$\delta(.) = 1$ if this argument is true and 0 otherwise}
```

Figure 8. Bagging algorithm

After training the classifiers, an observation is assigned to the class that has the majority vote. The bagged classifier has significantly greater accuracy than a single classifier. However, as it decreases the generalization error by reducing the variance of the base classifiers, the performance depends on the stability of the base classifiers. If a base classifier is unstable, by reducing the errors associated with random fluctuations in training data, the accuracy can be increased. On the other hand, if a base classifier is stable, the error of the bagging classifier is primarily caused by bias in the base classifier and bagging may not be able to improve the performance significantly, even it may decrease the performance as the training set used contains only the 63% of the original data set.

The other method for increasing the accuracy is boosting, which iteratively changes the distribution of observations in the training set to focus on observations that are hard to classify (Tan, Steinbach, & Kumar, 2006). There are several boosting algorithms, which differ in the way the weights are updated at the end of each boosting round and the way the predictions of each classifier are combined.

Adaboost is a popular boosting algorithm proposed by Freund and Schapire, where the importance of a base classifier (M_i) depends on its error rate. Figure 9 (Tan, Steinbach, & Kumar, 2006) represents the basic procedure of AdaBoost.

```
AdaBoost Algorithm
(1) w = {$w_j$ = 1/$N$ | j = 1,2,…, $N$}  // Initialize the weights for all $N$ observations.
(2) Let $D_i$ be a bootstrap sample, $k$ be the number of boosting rounds.
(3) for $i$ = 1 to $k$ do
(4)      Create a bootstrap sample $D_i$ according to w.
(5)      Train a base classifier $M_i$ on $D_i$.
(6)      Apply $M_i$ to all observations in the original training set, $D$.
(7)      $error(M_i) = \frac{1}{N}\left[\sum_{j=1}^{N} w_j \times \delta(M_i(x_j) \neq y_j)\right]$   // Calculate the weighted error.
(8)      if error($M_i$) > 0.5 then
(9)              w = {$w_j$ = 1/N | j = 1,2,…, N}.  // Reset the weights for all N observations.
(10)             Go to Step 4.
(11)     end if
(12)     $\alpha_i = \frac{1}{2}\ln\left(\frac{1 - error(M_i)}{error(M_i)}\right)$
(13)     Update the weight of each observation according to:
             $w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \times \begin{cases} \exp^{-\alpha_j} \; if \; M_j(x_i) = y_i, \\ \exp^{-\alpha_j} \; if \; M_j(x_i) \neq y_i \end{cases}$
(14) end for
(15) $M^*(x)$ = $\mathrm{argmax}_y \Sigma_i \delta(M_i(x) = y)$.
{$\delta$(.) = 1 if this argument is true and 0 otherwise}
```

Figure 9. AdaBoost boosting algorithm

Boosting changes the distribution by assigning a weight to each training observation. Training sets are formed by sampling with replacement regarding these weights. Hence, an observation can be selected more than once. In each boosting round, a classifier is learned on the selected training set and at the end weights are updated. Initial weight for all observations is *1/N* where *N* is the number of training observations. Thus each observation has the equal probability of being selected at the beginning. Afterwards, for the observations that are misclassified, the weights are increased and for those that are correctly classified, the weights are decreased and each subsequent classifier is trained on a set which contains observations with higher weights. At the end, boosting classifier takes a weighted average of the predictions made by each classifier as its final prediction.

Medical Data Mining

Knowledge management and data mining techniques are adopted in various successful biomedical applications in recent years. Knowledge management techniques and methodologies have been used to support the management of multimedia and mission critical tacit and explicit biomedical knowledge. Data mining techniques have been used to discover various biological, drug discovery, and clinical diagnosis and prognosis patterns using selected statistical analyses, machine learning and neural networks methods. (Chen, Fuller, Friedman, & Hersh, 2005).

Several data mining techniques from different research fields are used in biomedical applications (Chen, Fuller, Friedman, & Hersh, 2005). Statistical techniques such as regression analysis, discriminant analysis, time series analysis, principal component analysis, and multi-dimensional scaling are used widely as benchmarks for comparison with other techniques, such as Bayesian learning from pattern recognition. Bayesian learning has been widely used in biomedical data mining research, in particular, genomic and microarray analysis. Other techniques from machine learning are rule induction/decision trees and support vector machines (SVMs) that are also suitable for various biomedical classification problems, such as disease state classification based on genetic variables or medical diagnosis based on patient indicators. Artificial neural networks have been also used in experiments and critical biomedical classification and clustering problems due to their predictive power and classification accuracy.

The disease identification, diagnosis and the prediction of patient outcome, prognosis are the fundamental activities of medicine, which can be investigated using probabilistic models. The desirable features of a general clinical probability model are listed by Dybowski and Roberts (2005) as follows: high accuracy, namely predicting classes correctly, high discrimination or in other words the lowest possible misclassification rate, accessible interpretation that explains input-output relations instead of a black box model, short construction time if the model is being updated regularly with new data, short running time particularly for real-time applications, robust to missing data that are a common problem in medical data sets and ability to incorporate pre-existing knowledge in order to aid model development and ease the interpretation of a model.

According to Braithwaite, Dripps, Lyon, and Murray (2001), when developing a system for a medical application, when comparing the results of above mentioned techniques, the selectivity measure that represents the false alarm rate of the system is very important, especially when the clinical system requires low levels of false alarm. Besides, sensitivity is of extreme importance as it shows the number of actual occurrences of a condition remain undiagnosed.

CHAPTER 4

PROBLEM STATEMENT

This chapter defines the aim of this study, introduces the dataset and data preprocessing steps before applying classification techniques. All analysis in the study was generated using SAS Enterprise Miner® software, Version 5.3 of the SAS System for Unix.

Problem Definition

The increasing number of hospital infection incidents with considerable physical and moral costs causes worries in health care environment. The aim of this research project is to discover a pattern with prominent reasons of hospital infections on newborn by applying data mining techniques on the data collected in Eskişehir Osmangazi University Practice and Research Hospital.

The greatest challenge in handling the data analysis was similar to most other medical data problems such as small sample size, difficulties in measuring attributes, inconsistencies due to manual data collection and missing values. In order to overcome these problems, domain knowledge was investigated; expert opinion and related assumptions were taken into consideration.

KDD process was strictly followed in the study to avoid interesting but meaningless and for end-user useless results.

Data Description

The data is provided by Department of Clinical Microbiology and Infectious Diseases, Eskişehir Osmangazi University, Faculty of Medicine from the study to determine the prominent indicators of hospital infections and to prepare a form to track the progress of these indicators in newborn unit overall Turkey.

The data were collected between January 1, 2005 and December 31, 2005 in the NICU with a capacity of serving 16 patients. During the observation period 545 patients admitted to the unit and 120 of them were diagnosed with hospital infections. 102 attributes representing those patients were collected. Common characteristics of these attributes are related to patient, medical problem and treatment:

- information about the patient such as gestational age, weight, premature, congenital anomaly,
- indicators of particular medical problems such as meconium aspiration syndrome, acute renal failure, perinatal asphyxia, and
- indicators of particular medical treatments such as phototherapy, mechanical ventilation, catheter or antibiotic usage.

Variables such as lab-, antibiotics-, and hospital-costs, types of infections, types of bacteria causing the infection, C-reactive protein test results as an indicator of infection and exitus as the indicator of death were excluded in the study, as they were determined during/after the diagnosis. The list of 83 variables considered in the study is presented in Table 1. The dependent variable is *HospInfec*. If the newborn is infected, the variable takes the value of "1", else "0".

Table 1. List of Variables

Variable	Type	Explanation
AgeDay	Ratio	Age of the observation in days as of NBU admission
AntibioticUsage	Nominal (Binary)	Indicator of antibiotic usage
APGAR	Ordinal	APGAR score of the observation
ARF	Nominal (Binary)	Indicator of acute renal failure
Aspiration	Nominal (Binary)	Indicator of aspiration
AspirDur	Ratio	Aspiration duration
BirthOption	Nominal (Binary)	Indicator of birth option: c-section or normal
BirthPlace	Nominal (Binary)	Indicator of birth place: inside or outside a health center
BirthWeek	Ratio	Gestational age
BirthWGroup2	Ordinal (Binary)	Gestational age in 2 groups (0: BirthWeek =<32, 1: BirthWeek >32)
BirthWGroup4	Ordinal	Gestational age in 4 groups (1: BirthWeek =<28, 2: 28< BirthWeek =<32, 3: 32< BirthWeek =<36, 4: BirthWeek >36)
CatRelBloodFlow	Nominal (Binary)	Indicator of catheter-associated blood flow infection
CentCatDur	Ratio	Central Catheter Duration
CentCatheter	Nominal (Binary)	Indicator of centeral catheter usage
ChestTube	Nominal (Binary)	Indicator of chest tube
Chorioamnio	Nominal (Binary)	Indicator of chorioamnionitis
CongAno	Nominal (Binary)	Indicator of congenital anomaly
CutDown	Nominal (Binary)	Indicator of cut down
DetailBirthPlace	Nominal	Detailed Birth Place: Birth Center, Hospital, OGU Hospital, Home, Other
EMR	Nominal (Binary)	Indicator of early membrane rupture
ENDur	Ratio	Enteral nutrition duration in days
EnteralNut	Nominal (Binary)	Indicator of enteral nutrition
Gender	Nominal (Binary)	Gender of the observation (0: boy, 1: girl)
h2blockPPI	Nominal (Binary)	Indicator of H2 blockers or Proton-pump inhibitors (PPIs)
HBM	Nominal (Binary)	Indicator of hyperbilirubinemia
HospDur	Ratio	Hospital duration in days
HospDurLong2	Ordinal (Binary)	Indicator of long hospital duration, long if duration is greater than or equal to 7 days
HospInfec	Nominal (Binary)	Indicator of hospital infections

Table 1. Continued

Variable	Type	Explanation
HydropsFetalis	Nominal (Binary)	Indicator of hydrops fetalis
Immunosupp	Nominal (Binary)	Indicator of immunosuppression
InfectionDur	Ratio	Infection duration
ICU	Nominal (Binary)	Indicator of newborn intensive care unit (NICU) stay
ICUDur	Ratio	Duration in NICU
Intubation	Nominal (Binary)	Indicator of intubation
IntubDur	Ratio	Intubation duration
Invasive	Nominal (Binary)	Indicator of invasive procedures
IUGR	Nominal (Binary)	Indicator of intrauterine growth restriction
LBW	Nominal (Binary)	Indicator of low birth weight
LowAPGAR	Nominal (Binary)	Indicator of low APGAR score, low if score is less than 5
LowGISBleed	Nominal (Binary)	Indicator of low gastrointestinal system bleeding
MecAspSyn	Nominal (Binary)	Indicator of meconium aspiration syndrome
MechVent	Nominal (Binary)	Indicator of mechanical ventilation
NasogTube	Nominal (Binary)	Indicator of nasogastric intubation
Ncpap	Nominal (Binary)	Indicator of nasal continuous positive airway pressure (nCPAP)
NcpapDur	Ratio	Nasal continuous positive airway pressure (nCPAP) duration
NgUseDur	Ratio	Duration of nasogastric tube usage
NurseNum	Ratio	Number of nurses serving in NBU as of patient's admission
OgUseDur	Ratio	Duration of orogastric tube usage
Oligohyd	Nominal (Binary)	Indicator of oligohydramnios
OrogTube	Nominal (Binary)	Indicator of orogastric intubation
PAsphyxia	Nominal (Binary)	Indicator of perinatal asphyxia
PatientNum	Ratio	Number of patients in NBU as of patient's admission
PerCatDur	Ratio	Peripheral Catheter Duration
PeripCatheter	Nominal (Binary)	Indicator of peripheral catheter usage
Phlebotomy	Nominal (Binary)	Indicator of phlebotomy
PhototerDur	Ratio	Phototherapy duration
Phototherapy	Nominal (Binary)	Indicator of phototherapy

Table 1. Continued

Variable	Type	Explanation
Polycythemia	Nominal (Binary)	Indicator of polycythemia
PPV	Nominal (Binary)	Indicator of positive pressure ventilation
PPVDur	Ratio	Positive pressure ventilation duration
Premature	Nominal (Binary)	Indicator of premature birth
Preterm	Nominal (Binary)	Indicator of preterm parturition syndrome
RDS	Nominal (Binary)	Indicator of respiratory distress syndrome
SteroidUsage	Nominal (Binary)	Indicator of steroid usage
SurgInterv	Nominal (Binary)	Indicator of surgical intervention
TPN	Nominal (Binary)	Indicator of total parenteral nutrition
TPNDur	Ratio	Total parenteral nutrition duration
TPNLipid	Nominal (Binary)	Indicator of total parenteral feeding including lipids
Trakeos	Nominal (Binary)	Indicator of tracheostomy
TTN	Nominal (Binary)	Indicator of transient tachypnea of newborn
Twin	Nominal (Binary)	Indicator of twin sister/brother
UmbCatDur	Ratio	Umbilical Catheter Duration
UmbCatheter	Nominal (Binary)	Indicator of umbilical catheter usage
UriCatDur	Ratio	Urinary Catheter Duration
UriCatRelUrinary	Nominal (Binary)	Indicator of urinary catheter-associated urinary infection
UrinCatheter	Nominal (Binary)	Indicator of urinary catheter usage
VenDur	Ratio	Ventilation Duration
VenRelPneum	Nominal (Binary)	Indicator of ventilatory-associated pneumonia
Weight	Ratio	Weight of the observation at birth
WeightGroup2	Ordinal (Binary)	Birth weight in 2 groups (0: Weight > 2500, 1: Weight =< 2500)
WeightGroup3	Ordinal	Birth weight in 3 groups (1: Weight =<1500, 2: 1500< Weight =<2500, 3: 2500< Weight)
WeightGroup5	Ordinal	Birth weight in 5 groups (1: Weight =<1000, 2: 1000< Weight =<1500, 3: 1500< Weight =< 2000, 4: 2000< Weight =<2500, 5: Weight >2500)
WrappedCord	Nominal (Binary)	Indicator of wrapped cord

Variables that indicate either a treatment type or a device used during the treatment were structured in both nominal (binary) and ratio, such as phototherapy treatment (applied or not) and phototherapy treatment duration, or peripheral catheter usage (used or not) and peripheral catheter usage duration. Some ratio-scaled variables such as birth week and weight were used to create derived variables in ordinal scale, such as *BirthWeekGroup2* (binary) and *BirthWeekGroup4*, or *WeightGroup2* (binary), *WeightGroup3*, *WeightGroup4*.

Data Preprocessing

Before applying the classification algorithms, data were cleaned, important features were identified and data were transformed into appropriate forms. Dataset was checked for duplicate variables with correlation and chi-square analysis. No duplicates were detected.

Data Cleaning

According to descriptive data analysis; inconsistencies and incomplete variables were identified. Inconsistencies in the dataset that arose from manual coding mistakes were eliminated through data provider’s support and related assumptions. The most important assumption was to check the binary indicators with their associated duration variables: If the duration variable is greater than zero, then binary indicator is one, else zero. All binary variables having the same condition were checked with the rule and for 5 variables including intubation, peripheral catheter, urinary catheter, phototherapy and intensive care unit stay, necessary corrections were done. Similar check for group variables such as *HospDurLong2*, *WeightGroup2*, *WeightGroup3*, *WeightGroup5*, *BirthWGroup2*, *BirthWGroup4*, which were derived from other continuous variables, was also done and errors in these variables were corrected. Apart from these assumptions, *BirthPlace* variable, which indicates the birth inside/outside a health center, was inconsistent with *DetailBirthPlace* variable. Therefore, this variable was excluded and a new variable was defined as *BirthHealthCenter* that fulfills the definition of the existing variable. Table 2 represents an example for identified inconsistencies where the binary indicator of phototherapy treatment of five observations were assigned as "yes", whose phototherapy treatment duration were zero and phototherapy treatment of three observations were assigned as "no", whose phototherapy treatment duration were 3 and 14 days.

Table 2. Example for Inconsistency: Phototherapy Treatment vs. Duration

	PhototerDur																		
Phototherapy	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19
0	190	.	.	2	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
1	5	19	72	87	61	37	18	12	11	8	6	3	2	1	5	1	1	2	1

Before Data Cleaning

	PhototerDur																		
Phototherapy	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19
0	195	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	.	19	72	89	61	37	18	12	11	8	6	3	2	1	6	1	1	2	1

After Data Cleaning

After eliminating inconsistencies, in order to enrich the data, new variables were created such as weight-birth week ratio, in order to create a growth index for newborn infants. Other variables were created as ratios for ICU, Aspiration, Urinary Catheter, Peripheral Catheter, Enteral Nutrition, TPN, and Intubation durations, in order to identify the percentage of treatment duration to overall hospital stay. Variable that represents detail birth place was also used to derive binary variables for each category. Table 3 introduces the type and formula of each new derived variable.

Table 3. Derived Variables

New Variable	Type	Formula
WeightBWeekRat	Ratio	Weight / BirthWeek
ICUDurRat	Ratio	ICUDur / HospDur
AspirDurRat	Ratio	AspirDur / HospDur
UriCatDurRat	Ratio	UriCatDur / HospDur
PerCatDurRat	Ratio	PerCatDur / HospDur
ENDurRat	Ratio	ENDur / HospDur
TPNDurRat	Ratio	TPNDur / HospDur
IntubDurRat	Ratio	IntubDur / HospDur
BirthPlaceOGU	Nominal (Binary)	if DetailBirthPlace = 'OGU' then 1 else 0
BirthPlaceHosp	Nominal (Binary)	if DetailBirthPlace = 'Hospital' then 1 else 0
BirthPlaceBCent	Nominal (Binary)	if DetailBirthPlace = 'Birth Center' then 1 else 0
BirthPlaceHome	Nominal (Binary)	if DetailBirthPlace = 'Home' then 1 else 0
BirthPlaceOth	Nominal (Binary)	if DetailBirthPlace = 'Other' then 1 else 0
BirthHealthCenter	Nominal (Binary)	if DetailBirthPlace in ('OGU','Birth Center','Hospital') then 1 else 0

If any value of existing variables was missing, the value of new derived variable was also remained missing.

In order to identify the basic characteristics of the data set descriptive data analysis was conducted. Following tables Table 4 and Table 5 represent the descriptive statistics for continuous and discrete variables respectively.

Table 4. Descriptive Statistics for Continuous Variables

Variable	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum
AgeDay	2.54	4.889	545	0	0	1	46
AspirDur	1.16	6.714	544	1	0	0	102
AspirDurRat	0.04	0.133	544	1	0	0	1
BirthWeek	36.66	3.007	545	0	25	38	42
CentCatDur	0.01	0.3	545	0	0	0	7
ENDur	3.44	6.118	505	40	0	0	45
ENDurRat	0.34	0.429	505	40	0	0	1
HospDur	11.49	12.395	545	0	2	7	102
ICUDur	2.17	8.511	545	0	0	0	102
ICUDurRat	0.08	0.215	545	0	0	0	1
InfectionDur	0	0	321	224	0	0	0
IntubDur	1.59	6.995	545	0	0	0	102
IntubDurRat	0.07	0.194	545	0	0	0	1
NcpapDur	1.24	3.011	545	0	0	0	22
NgUseDur	1.21	6.456	545	0	0	0	72
NurseNum	2.00	0.066	227	318	2	2	3
OgUseDur	1.16	4.411	545	0	0	0	37
PatientNum	13.84	0.975	306	239	10	14	16
PerCatDur	9.1	12.274	545	0	0	5	102
PerCatDurRat	0.71	0.379	545	0	0	0.89	1
PhototherDur	2.77	3.178	545	0	0	2	19
PPVDur	0.12	0.455	545	0	0	0	5
TPNDur	4.72	11.307	545	0	0	0	102
TPNDurRat	0.19	0.313	545	0	0	0	1
UmbCatDur	0.12	0.564	545	0	0	0	7
UriCatDur	0.13	1.508	545	0	0	0	30
UriCatDurRat	0.01	0.06	545	0	0	0	1
VenDur	1.28	6.784	545	0	0	0	102
Weight	2734.99	800.807	545	0	400	2850	4700
WeightBWeekRat	73.54	17.836	545	0	13.333	75.14	120.51

For discrete variables, cross tabulations were prepared which can be found in Appendix B. Following Table 5 represents chi-square and p-value for each discrete variable.

Table 5. Descriptive Statistics for Discrete Variables

Input	Chi-Square	P-Value
AntibioticUsage	26.911185	<0.0001
APGAR	13.346331	0.2049
ARF	0.0365955	0.8483
Aspiration	20.524689	<0.0001
BirthHealthCenter	0.6340545	0.4259
BirthOption	0.021441	0.8836
BirthPlace	13.873498	0.0002
BirthPlaceBCent	4.0763363	0.0435
BirthPlaceHome	2.0021867	0.1571
BirthPlaceHosp	2.4926842	0.1144
BirthPlaceOGU	6.6093551	0.0101
BirthWGroup2	12.538471	0.0004
BirthWGroup4	16.732231	0.0008
ChestTube	0.3779645	0.5387
CongAno	1.85E-04	0.9891
DetailBirthPlace	11.542171	0.0211
EMR	0.4051757	0.5244
EnteralNut	17.308259	<0.0001
Gender	0.8463565	0.3576
h2blockPPI	10.243074	0.0014
HBM	1.844177	0.1745
HospDurLong2	24.515363	<0.0001
ICU	9.295349	0.0023
Intubation	20.969687	<0.0001
Invasive	0.8261434	0.3634
IUGR	0.0314932	0.8591
LBW	0.3672531	0.8322
LowAPGAR	0.1159999	0.7334
MecAspSyn	1.16E-04	0.9914
MechVent	20.315967	<0.0001
NasogTube	16.983699	<0.0001
Ncpap	8.3718351	0.0038
Oligohyd	0.3779645	0.5387
OrogTube	16.37782	<0.0001
PAsphyxia	5.311544	0.0212
PeripCatheter	18.811749	<0.0001

Table 5. Continued

Input	Chi-Square	P-Value
Phototherapy	0.4008332	0.5267
Polycythemia	1.8011659	0.1796
PPV	2.0425357	0.153
Premature	5.2256219	0.0223
Preterm	3.2915096	0.0696
RDS	8.3421977	0.0039
SteroidUsage	14.477049	0.0001
SurgInterv	16.568147	<0.0001
TPN	10.92501	0.0009
TPNLipid	10.567862	0.0012
TTN	0.5295585	0.4668
Twin	0.1599636	0.6892
UmbCatheter	0.0092188	0.9235
UrinCatheter	41.164504	<0.0001
VenRelPneum	14.238025	0.0002
WeightGroup22	3.1516744	0.0758
WeightGroup3	7.0671625	0.0292
WeightGroup5	7.466892	0.1132
WrappedCord	0.9502869	0.3296

Missing values were handled differently for different classification techniques. Though, variables with a high percentage of missing values were excluded from the study at the beginning, such as variables *PatientNum* and *NurseNum*, indicating the number of patients and nurses in NICU as of observation's admission, while each had a 43.9% of missing.

For logistic regression and neural networks, missing values of all types of variables were handled with decision tree induction, where replacement values were estimated by analyzing each incomplete variable as a dependent variable, and the remaining variables were used as independent variables. Table 6 represents the variables, their new names used in the model and the number of missing values that were replaced with the estimates of decision tree induction.

Table 6. Handling Missing Values with Decision Tree Induction

Variable Name	Impute Method	Imputed Variable	Variable Type	Number of Missing	Percentage of missing
APGAR	Tree	IMP_APGAR	Ordinal	106	19.45
AspirDur	Tree	IMP_AspirDur	Interval	1	0.18
AspirDurRat	Tree	IMP_AspirDurRat	Interval	1	0.18
ENDur	Tree	IMP_ENDur	Interval	40	7.34
ENDurRat	Tree	IMP_ENDurRat	Interval	40	7.34
LowAPGAR	Tree	IMP_LowAPGAR	Binary	106	19.45
VenRelPneum	Tree	IMP_VenRelPneum	Binary	1	0.18

For different decision tree classification algorithms, missing values were either used in search for a split or they were assigned to the largest branch.

When using missing values in search for a split on an input, all observations with missing values were assigned to the same branch. The branch might or might not contain other observations. Missing values were treated as having the same unknown non-missing value for continuous variables and as a separate category for categorical variables. Thus the worth of split was computed with the same number of observations for each dependent variable and an association of the missing values with the values of independent variable could contribute to the predictive ability of the split (SAS Help and Documentation, n.d.).

Data Reduction

Constant variables (variables with a single value) were checked and tracheostomy indicator (*Trakeos*) was eliminated, as this method was not applied on any newborn. Central catheter duration variable (*CentCatDur*) whose distribution was too narrow was also eliminated from all models. In order to reduce the number of variables that were used in different models, heuristic feature subset selection methods were used.

Before decision trees no data reduction technique was used, while decision tree induction can already be used for data reduction (variable selection). For logistic regression, from variables that represent the same feature, the one with more distinct values were selected except the variables *UriCatDur* and *UrinCatheter*. Instead of *UriCatDur* whose distribution was too narrow, nominal (binary) variable *UrinCatheter* was selected. Thus following variables were eliminated from logistic regression model: *Aspiration*, *BirthWGroup2*, *BirthWGroup4*, *CentCatheter*, *EnteralNut*, *HospDurLong2*, *ICU*, *Intubation*, *NasogTube*, *Ncpap*, *OrogTube*, *PPV*, *PeripCatheter*, *Phototherapy*, *TPN*, *TPNLipid*, *UmbCatheter*, *UriCatDur*, *WeightGroup2*, *WeightGroup3*, *WeightGroup5*.

Neural networks are affected by the number of input variables in two different manners. First, as the number of input variables increases, the size of the network becomes large. This increases the overfitting risk and needs more training data. Second, complex networks take a long time to converge weights. Because of these two aspects, the variables that were selected by logistic regression model were used as the input variables in neural networks model.

Data Transformation

Variables were transformed with different methods before applying different classification techniques. Although it is not necessary to transform the data before logistic regression and decision trees, both techniques benefit from the transformation. Transformation methods used in the study are standardization, equi-width binning, and equalize spread by target variable. “Standardization” is the z-score normalization and it was tested before logistic regression and neural networks; however the resulting generalization error of the models were higher than the models with other transformations.

Before all decision trees except CART, equi-width binning method was used in order to eliminate the effect of outliers and noise and to prevent overfitting. By "Equi-width binning" the data values are grouped into *N* equally spaced interval based on the difference between the maximum and the minimum values. The size of bins *d* is determined by the following formula:

$$d = \frac{X_{\max} - X_{\min}}{N}$$

where $X_{\max}$ and $X_{\min}$ are the maximum and minimum values of variable *X*, *N* is the user defined number of bins. The continuous input variables of CHAID and entropy models were transformed into 4 bins.

Before neural networks, "equalize spread by target variable" method, which is a subset of Box-Cox transformations (Box & Cox, 1964), was used. As in Box-Cox transformations, there are two steps for transformation. Variables are first scaled to [0,1] with the following formula:

$$x' = \frac{\max((x - x_{\min}),0)}{x_{\max} - x_{\min}}$$

where *x* is the variable to be transformed and x' is the scaled variable. Then one of the following transformations, which has the smallest variance of the variances between target levels (*HospInfec* = 1 and *HospInfec* = 0) is selected.

$$x', \ln(x'), \mathrm{sqrt}(x'), e^{(x')}, (x')^{1/4}, (x')^2, (x')^4$$

Thus the variance in the interval variables between different levels of target is stabilized. Following Table 7 shows the transformations done with equalize spread by target variable and transformation statistics. In order to prevent undefined results, “1” is added to the scaled variable before logarithmic transformation.

Table 7. Equalize Spread Transformation for Neural Networks

Method	Variable Name	Formula	Min	Max	Mean	Std. Deviation	Skewness	Kurtosis
Original	IMP_ENDur		0	45	3.768999	6.423006	2.807749	9.457189
Original	PerCatDur		0	102	9.104587	12.27406	3.494914	16.93789
Original	UriCatDurRat		0	1	0.00679	0.060223	12.631394	179.4945
Computed	LOG_IMP_ENDur	log(_VAR_ + 1)	0	0.69	0.073272	0.114516	2.297406	5.970689
Computed	LOG_PerCatDur	log(_VAR_ + 1)	0	0.69	0.080487	0.095765	2.727292	10.09333
Computed	SQRT_UriCatDurRat	Sqrt(_VAR_)	0	1	0.012347	0.081546	8.158494	75.85001

Variables enteral nutrition duration and peripheral catheter duration were transformed using logarithm, variable urinary catheter duration ratio was transformed with squared root method. Transformation methods helped to decrease the generalization error and to increase the accuracy.

CHAPTER 5

EVALUATION OF MODEL RESULTS

The surveillance of hospital infections in newborn ICUs can be supported through detecting the risky patient profiles. For this purpose, decision trees, neural networks and logistic regression classification techniques explained in Chapter 3 were applied on hospital infections data set obtained from Department of Clinical Microbiology and Infectious Diseases of Eskişehir Osmangazi University Medical School. In this chapter, the results and the comparison of the models are discussed.

Approach for Building Models

In the study, five different models were built with three different classification techniques. These are CHAID, CART, and entropy decision tree models, neural networks model and logistic regression model. Models try to predict the probability of a newborn being infected by hospital infections.

For different models, different accuracy estimation and increasing methods were applied. First accuracy estimation method was selected to be holdout stratified sampling with 70% of training and 30% of test samples. This method was applied on decision trees and neural networks. Logistic regression model was built on the whole data set. The parameters of classification algorithms decided with this method were used in other accuracy estimation and increasing methods. This approach facilitated parameter selection and detailed analyses of model results. Following Table 8 illustrates the frequency distribution of dependent variable in the whole sample and training and test samples.

Table 8. Whole Sample, Training and Test Samples

Dataset	Variable	Value	Frequency Count	Percentage
Whole	*HospInfec*	0	425	77.98%
	HospInfec	1	120	22.02%
Training	*HospInfec*	0	127	77.44%
	HospInfec	1	37	22.56%
Test	*HospInfec*	0	298	78.22%
	HospInfec	1	83	21.78%

Second accuracy estimation method was 10-fold cross validation. The data set was randomly partitioned into 10 folds and generalization error of the applied algorithm was estimated. The method was applied on each model and the comparison of the models was done according to the test sample results of this method.

Moreover, bagging and boosting accuracy increasing methods were conducted on each decision tree and neural network. For bagging models 10 random samples were created with replacement. Boosting models were created with different number of iterations for each model.

In order to assess the goodness of fit, average squared error, area under ROC curve, gini coefficient, error rate, accuracy, sensitivity and specificity measures were calculated. In addition, ROC curves were drawn for each model.

In the following sections the results of all models with applied accuracy estimation and accuracy increasing methods and the model comparison are given.

Decision Trees

In the study, three types of decision tree algorithms were conducted: CHAID, CART and customized decision trees using entropy splitting criterion. From accuracy estimation techniques, holdout and cross validation, from accuracy increasing techniques bagging and boosting were applied. There are 89 input variables for decision trees. These are listed in Table 9.

Table 9. Input Variables of Decision Trees

Input Variables		
APGAR	HospDur	Phlebotomy
ARF	HospDurLong2	PhototherDur
AgeDay	HydropsFetalis	Phototherapy
AntibioticUsage	ICU	Polycythemia
AspirDur	ICUDur	Premature
AspirDurRat	ICUDurRat	Preterm
Aspiration	IUGR	RDS
BirthHealthCenter	ImmunoSupp	SteroidUsage
BirthOption	IntubDur	SurgInterv
BirthPlaceBCent	IntubDurRat	TPN
BirthPlaceHome	Intubation	TPNDur
BirthPlaceHosp	Invasive	TPNDurRat
BirthPlaceOGU	LBW	TPNLipid
BirthPlaceOth	LowAPGAR	TTN
BirthWGroup2	LowGISBleed	Twin
BirthWGroup4	MecAspSyn	UmbCatDur
BirthWeek	MechVent	UmbCatheter
CatRelBloodFlow	NasogTube	UriCatDur
CentCatheter	Ncpap	UriCatDurRat
ChestTube	NcpapDur	UriCatRelUrinary
Chorioamnio	NgUseDur	UrinCatheter
CongAno	OgUseDur	VenDur
CutDown	Oligohyd	VenRelPneum
EMR	OrogTube	Weight

Table 9. Continued

Input Variables		
ENDur	PAsphyxia	WeightBWeekRat
ENDurRat	PPV	WeightGroup22
EnteralNut	PPVDur	WeightGroup3
Gender	PerCatDur	WeightGroup5
h2blockPPI	PerCatDurRat	WrappedCord
HBM	PeripCatheter	

Before CHAID and entropy decision tree models, all continuous variables were transformed with equi-width binning where the variables were binned into 4 groups.

CHAID Decision Tree Model

In the study, following parameters illustrated in Table 10 were set in order to build the CHAID decision tree model. Maximum tree depth, minimum leaf size and minimum categorical size parameters were used as stopping rules for the decision tree growth. Minimum categorical size parameter indicates the number of observations that a categorical value must have before the category can be used in a split search.

Table 10. CHAID Decision Tree Parameters

Parameter	Value
Significance Level for Split	0.05
Significance Level for Merge	0.05
Maximum Tree Depth	6
Minimum Leaf Size	5
Minimum Categorical Size	5

CHAID Decision Tree Model with Holdout Sampling

CHAID decision tree model selected *AntibioticUsage, ENDur, PerCatDur,* and

UrinCatheter variables with holdout sampling. The figure of tree structure can be found in Appendix C, Figure 34 and rules derived from the tree are illustrated in following Figure 10.

```
IF  UrinCatheter EQUALS 1
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N    :     7
 0    :  28.6%
 1    :  71.4%

IF  Transformed: ENDur EQUALS 01:LOW -11.25
AND Transformed: PerCatDur EQUALS 02:25.5-51 THEN
 N    :     5
 0    :  80.0%
 1    :  20.0%

IF  Transformed: ENDur EQUALS 02:11.25-22.5
AND Transformed: PerCatDur EQUALS 02:25.5-51 THEN
 N    :     7
 0    :  28.6%
 1    :  71.4%

IF  Transformed: ENDur EQUALS 03:22.5-33.75
AND Transformed: PerCatDur EQUALS 02:25.5-51 THEN
 N    :    16
 0    :   0.0%
 1    : 100.0%

IF  AntibioticUsage EQUALS 0 AND UrinCatheter EQUALS 0
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N    :   127
 0    :  92.9%
 1    :   7.1%

IF  AntibioticUsage EQUALS 1 AND UrinCatheter EQUALS 0
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N    :   219
 0    :  78.5%
 1    :  21.5%
```

Figure 10. Rules of CHAID decision tree model

All variables of the model are related to medical treatments. Analyzing the effects of the variables, it is seen that as enteral nutrition (*ENDur*) and peripheral catheter durations (*PerCatDur*) increase, the probability of hospital infections increases. In addition, the usage of antibiotics and urinary catheters increases the risk of hospital infections. The effects of these variables on hospital infections are consistent with the medical literature mentioned in Chapter 2. Figure 10 shows the rules represented by CHAID decision tree model. In the figure, N is the number of observations in a given node. The percentage of “1” represents the probability of infected and “0” represents the probability of noninfected newborns.

The goodness of fit statistics and ROC curves for training and test samples are given in Table 11 and Figure 11 respectively.

Table 11. CHAID Holdout Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.13	0.15
Area under ROC	0.73	0.68
Gini Coefficient	0.47	0.36
Error Rate	0.16	0.18
Accuracy	0.84	0.82
Sensitivity	0.31	0.30
Specificity	0.99	0.98

According to the differences in the statistics and ROC curves between training and test samples, it can be claimed that CHAID decision tree model successfully handled the risk of overfitting.

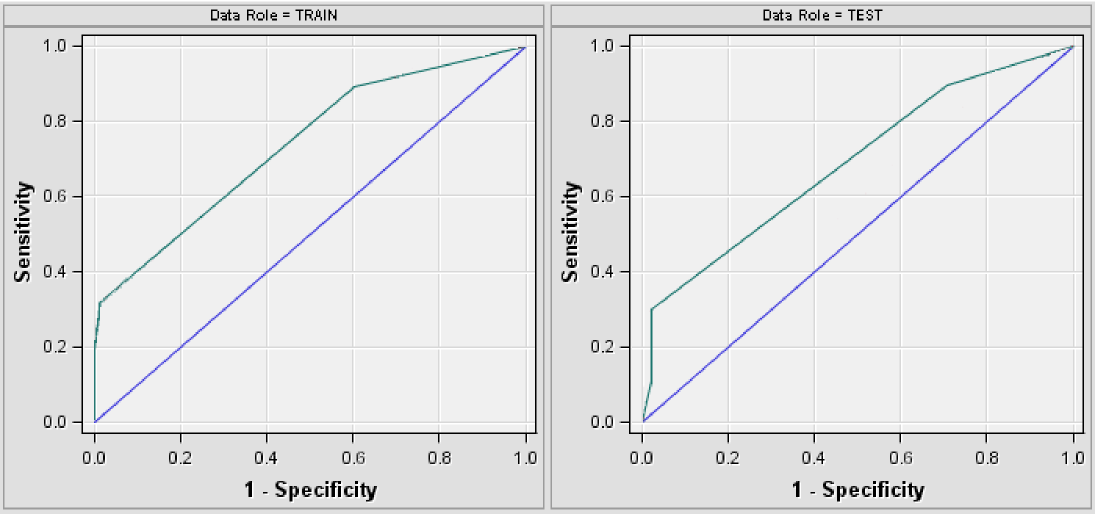

Figure 11. CHAID holdout model ROC curves

However, sensitivity and specificity measures indicate that the model poorly performs on infected newborns, whereas it performs considerably well on noninfected ones.

CHAID Decision Tree Model with Cross Validation

Variables *AntibioticUsage, ENDurRat, EnteralNut, HospDurLong2, LowAPGAR, PerCatDur, TPNDur,* and *UrinCatheter* were selected by CHAID models built with cross validation method. In addition to the variables related to medical treatments, patient characteristic low APGAR score indicator was also selected by one or more models.

The model results represented in Table 12 shows that area under ROC curve, gini coefficient and sensitivity were decreased in test sample. Still, the error rate is stable in training and test samples and also close to the error rate of holdout sampling.

Table 12. CHAID Cross Validation Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.14	0.14
Area under ROC	0.73	0.68
Gini Coefficient	0.45	0.36
Error Rate	0.17	0.17
Accuracy	0.83	0.83
Sensitivity	0.33	0.30
Specificity	0.97	0.98

As it was in holdout sampling, sensitivity measure of cross validation models indicates that the model has a poor performance on infected, but substantial performance on noninfected newborns. Figure 12 illustrates the ROC curve provided by the training folds in cross validation.

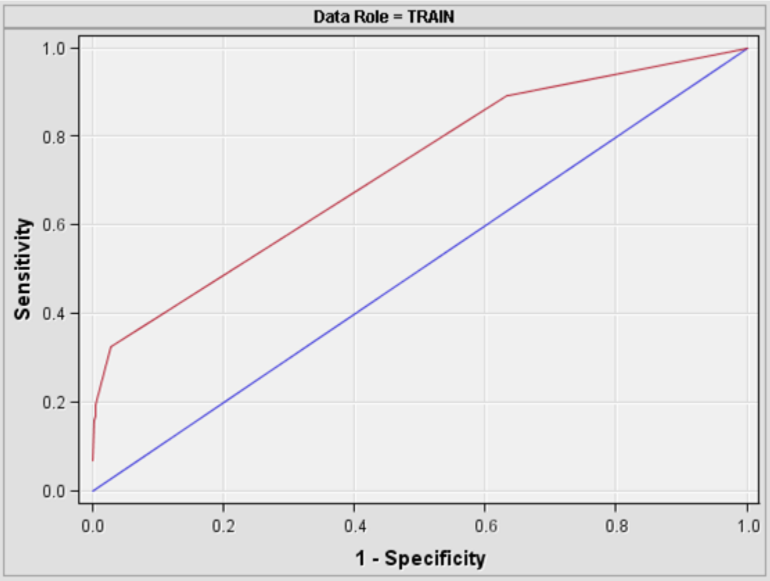

Figure 12. CHAID cross validation model ROC curve

CHAID Decision Tree Model with Bagging

In bagging method, 10 random samples were created with replacement. CHAID models were built on each of these samples. The goodness of fit statistics provided by training data set is demonstrated in Table 13.

Table 13. CHAID Bagging Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.14
Area under ROC	0.76
Gini Coefficient	0.52
Error Rate	0.18
Accuracy	0.82
Sensitivity	0.38
Specificity	0.95

Compared to the results of cross validation, bagging increased area under ROC curve, gini coefficient and sensitivity measures while a slight decrease was observed in accuracy and specificity. ROC curve provided by the training samples is illustrated in following Figure 13.

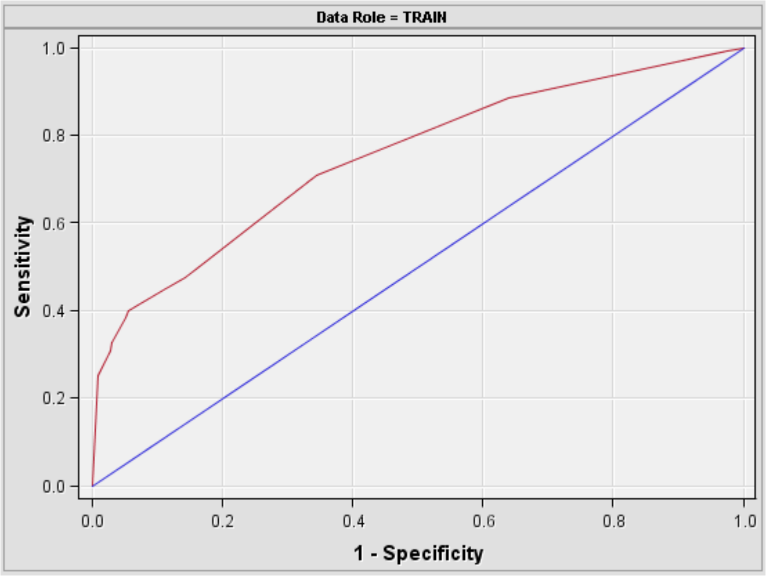

Figure 13. CHAID bagging model ROC curve

Bagging model ROC curve shows an improvement in upper left direction caused by the increase in sensitivity measure when compared to cross validation and holdout models.

CHAID Decision Tree Model with Boosting

Boosting method was conducted on CHAID decision trees with 20 iterations. In each iteration the training sample was formed by sampling with replacement regarding the weights assigned to each observation.

Table 14. CHAID Boosting Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.22
Area under ROC	0.74
Gini Coefficient	0.49
Error Rate	0.18
Accuracy	0.82
Sensitivity	0.24
Specificity	0.98

In comparison to cross validation training sample results 8% increase in average squared error and ca.10% decrease in sensitivity were observed, though the error rate and the accuracy did not change.

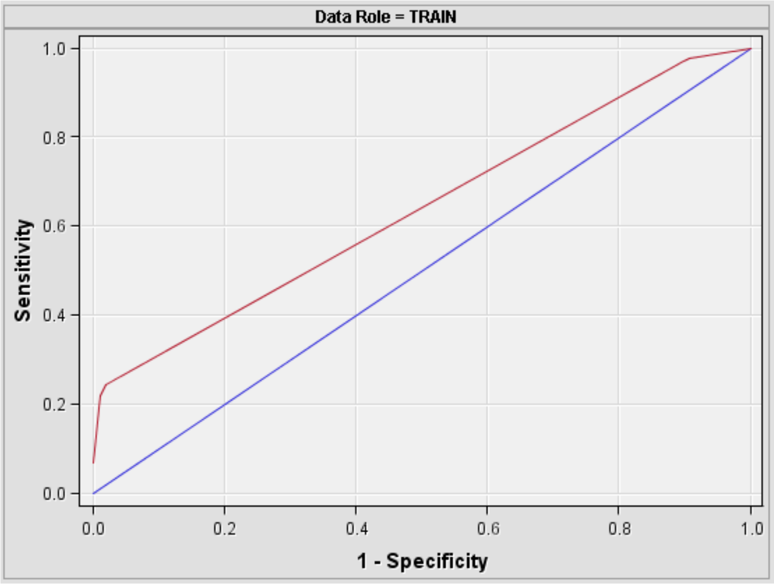

Figure 14. CHAID boosting model ROC curve

ROC curve of the model demonstrated in Figure 14 is more flat because of the decrease in sensitivity measure in comparison to cross validation method. Contrary to expectations, these results showed that boosting did not improve the model accuracy.

CART Decision Tree Model

CART Decision Tree model uses gini splitting criterion and creates binary decision trees. For CART algorithm applied in the study, the input variables are either nominal or ratio. Ordinal inputs are treated as interval. Therefore, no binning transformation was performed before applying CART models.

Missing values were handled by surrogate splits in the applied CART decision tree algorithm. Surrogate splits were created and used to assign observations to branches when the primary splitting variable was missing. If missing values could not be handled by surrogate rules, then the observation was assigned to the largest branch.

Moreover, the applied CART algorithm needed validation data set in order to realize cost-complexity pruning by comparing the average squared error between training and validation samples. As the data set was too small to separate into three partitions, for each subtree 10-fold cross validation method was conducted.

In addition to pruning methods, the tree growth was restricted with the stopping rules shown in Table 15.

Table 15. CART Decision Tree Parameters

Parameter	Value
Maximum Tree Depth	6
Minimum Leaf Size	5
Minimum Categorical Size	5

<u>CART Decision Tree Model with Holdout Sampling</u>

Variables *TPNDur* and *PerCatDur* were selected by CART decision tree model with holdout sampling. The figure of tree structure can be found in Appendix C, Figure 35 and rules derived from the tree are illustrated in following Figure 15. In figure, N is the number of observations in the node, and the percentages of “1” and “0” represent the probability of infected and noninfected newborns respectively.

```
IF 24.5 <= TPNDur  THEN
 N     :    20
 0     :  10.0%
 1     :  90.0%

IF PerCatDur < 4.5 AND TPNDur < 24.5  THEN
 N     :   138
 0     :  96.4%
 1     :   3.6%

IF 4.5 <= PerCatDur AND TPNDur < 24.5 THEN
 N     :   223
 0     :  73.1%
 1     :  26.9%
```

Figure 15. Rules of CART decision tree model

According to the classification rules represented by model, the probability of hospital infections increases, as total parenteral nutrition duration (*TPNDur*) and peripheral catheter duration (*PerCatDur*) variables increase. Both variables are related with medical treatments.

The goodness of fit statistics and ROC curves for training and test samples are given in Table 16 and Figure 16 respectively. The differences between training and test sample statistics are relatively high and affirm that the model is not successful on test as it is on training sample.

Table 16. CART Holdout Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.13	0.19
Area under ROC	0.75	0.56
Gini Coefficient	0.50	0.12
Error Rate	0.18	0.23
Accuracy	0.82	0.77
Sensitivity	0.22	0.11
Specificity	0.99	0.97

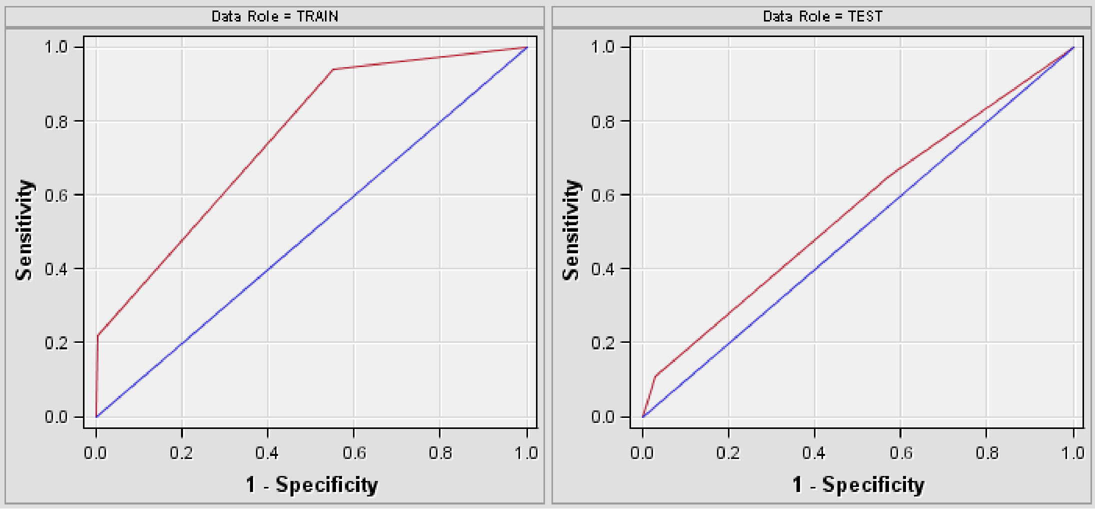

Figure 16. CART holdout model ROC curve

Sensitivity measure of the model is substantially low, where specificity measure indicates that the model performs well on noninfected newborns.

CART Decision Tree Model with Cross Validation

In CART decision tree model with cross validation 7 variables were selected by models. These are *AspirDurRat, ENDurRat, PerCatDur, PerCatDurRat, TPNDur, UriCatDur,* and *Weight*. However, because of surrogate rules, 23 additional variables were selected in order to handle missing values: *APGAR, AgeDay, AntibioticUsage, AspirDur, Aspiration, BirthWeek, ENDur, EnteralNut, HospDur, ICUDur, IntubDur, IntubDurRat, LBW, NcpapDur, NgUseDur, OgUseDur, PeripCatheter, TPNDurRat, UmbCatDur, UriCatDurRat, UrinCatheter,* and *WeightBWeekRat*. Most of the variables are related to medical treatments. In addition, patient characteristics *APGAR, AgeDay, BirthWeek, LBW*, *Weight* and *WeightBWeekRat* were also selected.

Goodness of fit statistics provided by averaging the training and test folds are represented in Table 17. In comparison to holdout model, cross validation results in training and test samples are more consistent.

Table 17. CART Cross Validation Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.14	0.15
Area under ROC	0.66	0.65
Gini Coefficient	0.32	0.31
Error Rate	0.17	0.19
Accuracy	0.83	0.81
Sensitivity	0.28	0.26
Specificity	0.99	0.97

The ROC curve of overall training folds in cross validation method is illustrated in Figure 17. Figure shows an improvement in sensitivity measure at lower values of "1-Specificity" in comparison to holdout model.

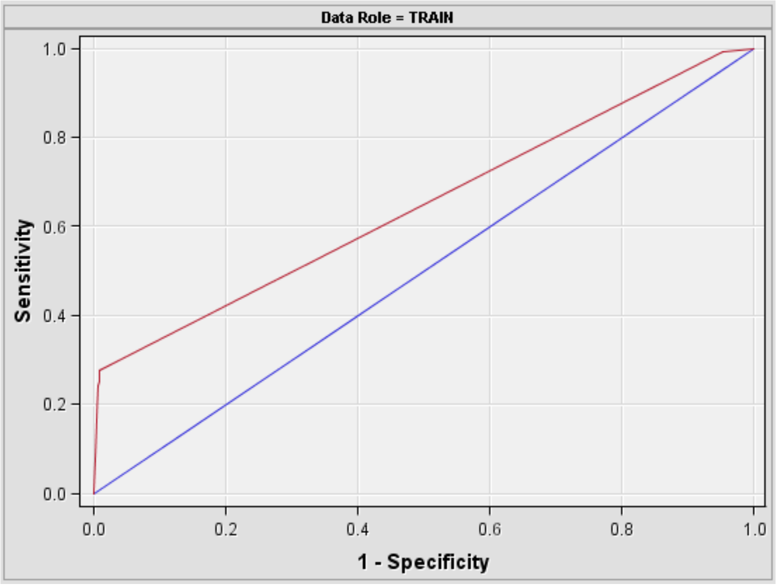

Figure 17. CART cross validation model ROC curve

CART Decision Tree Model with Bagging

As the base classifier, CART decision tree algorithm was applied on 10 training samples which were created with replacement from the whole data set. The results of bagged CART decision tree are demonstrated in Table 18.

Table 18. CART Bagging Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.14
Area under ROC	0.65
Gini Coefficient	0.29
Error Rate	0.17
Accuracy	0.83
Sensitivity	0.32
Specificity	0.97

When compared to cross validation method, the error rate of bagged CART decision tree was not changed. Though, the sensitivity and specificity measures were affected.

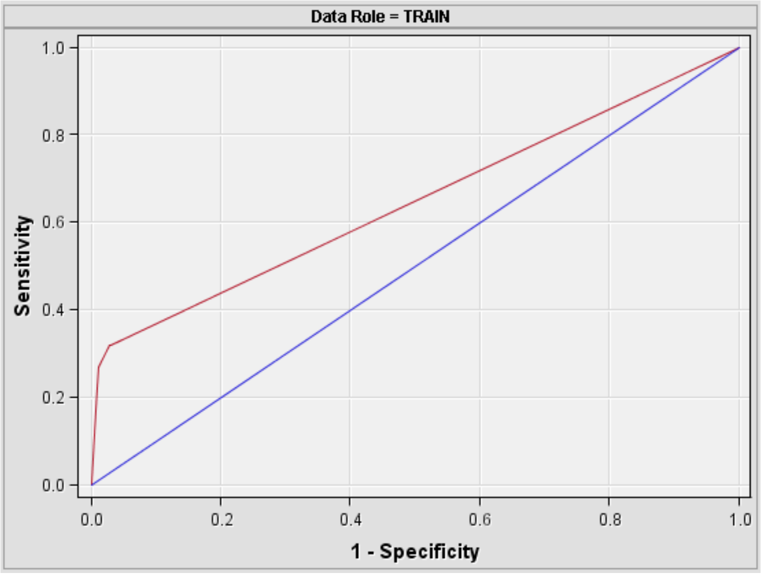

Figure 18. CART bagging model ROC curve

As seen in Figure 18, the increase in sensitivity and the decrease in specificity measures led to a slight improvement in the ROC curve.

<u>CART Decision Tree Model with Boosting</u>

Boosting was applied on CART decision trees with 10 iterations. However, the method failed because the error of CART model never fell below 0.5 through the iterations. As a result, the specificity reached to 1 and sensitivity to 0.

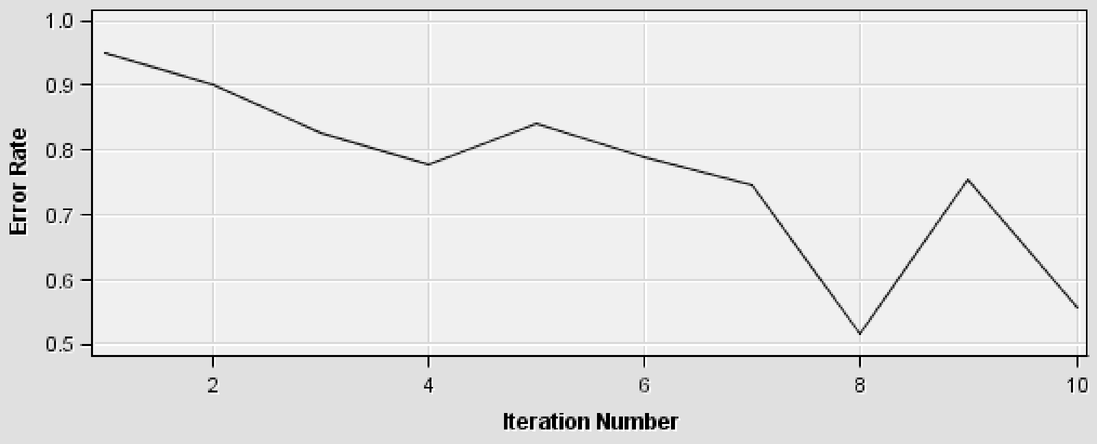

Figure 19. CART boosting iteration error

Entropy Decision Tree Model

Entropy decision tree models use entropy reduction (information gain) as splitting criterion. Missing values were used in search as described in Data Preprocessing section. The entropy decision trees were not pruned however, stopping rules were defined. Table 19 represents the parameters of entropy decision tree model.

Table 19. Entropy Decision Tree Parameters

Parameter	Value
Maximum Branch	5
Maximum Tree Depth	4
Minimum Leaf Size	10
Minimum Categorical Size	5

Entropy Decision Tree Model with Holdout Sampling

Entropy decision tree model with holdout sampling selected *AntibioticUsage, APGAR, ENDur, ENDurRat, Gender, HospDurLong2, Ncpap, PerCatDur* and *WeightBWeekRat* variables. Six variables including *AntibioticUsage, ENDur, ENDurRat, HospDurLong2, Ncpap, PerCatDur* were related to medical treatments, whereas other variables *APGAR, Gender* and *WeightBWeekRat* were related to patient characteristics.

The figure of entropy decision tree structure can be found in Appendix C, Figure 36 and the classification rules presented by the tree are illustrated in following Figure 21.

```
IF  APGAR IS ONE OF: 0 2 3 4 AND Transformed: PerCatDur EQUALS 01:LOW -25.5
THEN
 N      :     10
 1      :   0.0%
 0      :  100.0%

IF  APGAR EQUALS 5 AND Transformed: PerCatDur EQUALS 01:LOW -25.5  THEN
 N      :     10
 1      :  50.0%
 0      :  50.0%

IF  Transformed: ENDur IS ONE OF: 01:LOW -11.25 02:11.25-22.5
AND Transformed: PerCatDur EQUALS 02:25.5-51 THEN
 N      :     12
 1      :  50.0%
 0      :  50.0%

IF  Transformed: ENDur EQUALS 03:22.5-33.75
AND Transformed: PerCatDur EQUALS 02:25.5-51 THEN
 N      :     16
 1      :  100.0%
 0      :   0.0%

IF  Gender EQUALS 0 AND APGAR EQUALS 10
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N      :     12
 1      :   0.0%
 0      :  100.0%

IF  Gender EQUALS 1 AND APGAR EQUALS 10
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N      :     10
 1      :  10.0%
 0      :  90.0%

IF  Transformed: WeightBWeekRat EQUALS 01:LOW -53.378205128
AND HospDurLong2 EQUALS 1 AND APGAR IS ONE OF: 6 7 8
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N      :     27
 1      :  14.8%
 0      :  85.2%

IF  Transformed: WeightBWeekRat EQUALS 02:53.378205128-75.756410255
AND HospDurLong2 EQUALS 1 AND APGAR IS ONE OF: 6 7 8
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N      :     32
 1      :  15.6%
 0      :  84.4%
```

Figure 20. Rules of entropy decision tree model

```
IF  Transformed: WeightBWeekRat EQUALS 03:75.756410255-98.134615383
AND HospDurLong2 EQUALS 1 AND APGAR IS ONE OF: 6 7 8
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N     :    29
 1     :  37.9%
 0     :  62.1%

IF  Transformed ENDurRat EQUALS 04:0.75-HIGH AND HospDurLong2 EQUALS 0
AND APGAR IS ONE OF: 6 7 8 AND Transformed: PerCatDur EQUALS 01:LOW -25.5
THEN
 N     :    14
 1     :  21.4%
 0     :  78.6%

IF  Transformed ENDurRat EQUALS 01:LOW -0.25 AND HospDurLong2 EQUALS 0
AND APGAR IS ONE OF: 6 7 8 AND Transformed: PerCatDur EQUALS 01:LOW -25.5
THEN
 N     :    95
 1     :   3.2%
 0     :  96.8%

IF  Transformed ENDurRat EQUALS 04:0.75-HIGH AND AntibioticUsage EQUALS 0
AND APGAR EQUALS 9 AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N     :    29
 1     :  20.7%
 0     :  79.3%

IF  Transformed ENDurRat EQUALS 01:LOW -0.25 AND AntibioticUsage EQUALS 0
AND APGAR EQUALS 9  AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N     :    16
 1     :   0.0%
 0     : 100.0%

IF  Ncpap EQUALS 1 AND AntibioticUsage EQUALS 1 AND APGAR EQUALS 9
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N     :    22
 1     :  22.7%
 0     :  77.3%

IF  Ncpap EQUALS 0 AND AntibioticUsage EQUALS 1 AND APGAR EQUALS 9
AND Transformed: PerCatDur EQUALS 01:LOW -25.5 THEN
 N     :    47
 1     :  38.3%
 0     :  61.7%
```

Figure 20. Continued

According to the classification rules, it is seen that the increase in variables *PerCatDur*, *ENDur, ENDurRat* and *WeightBWeekRat* increases the risk of hospital infections. Moreover, antibiotics usage, long hospital stay, APGAR score between 5 and 9, newborn female indicate high risk of hospital infections. Variable *Ncpap* indicating the usage of nasal continuous positive airway pressure (nCPAP) was not found as a risk factor of hospital infections instead, usage of the method slightly contributes to find healthy newborns.

Table 20 represents the goodness of fit statistics and Figure 21 illustrates the ROC curves of entropy decision trees by training and test samples.

Table 20. Entropy Holdout Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.12	0.18
Area under ROC	0.83	0.67
Gini Coefficient	0.66	0.34
Error Rate	0.18	0.26
Accuracy	0.82	0.74
Sensitivity	0.33	0.16
Specificity	0.96	0.91

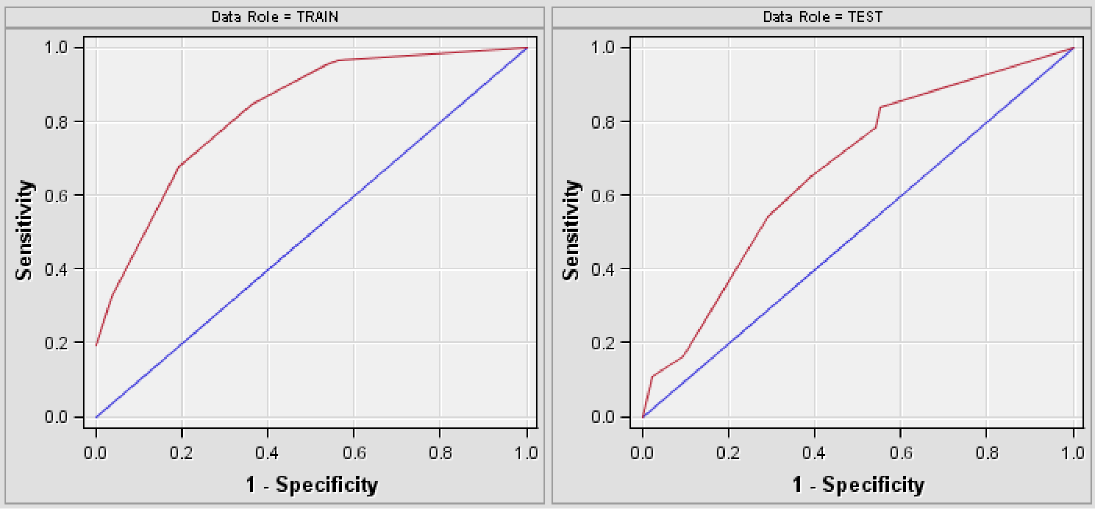

Figure 21. Entropy holdout model ROC curve

According to the results represented in goodness of fit statistics and ROC curves, it can be referred to the presence of overfitting problem.

Entropy Decision Tree Model with Cross Validation

Entropy decision tree models within cross validation folds selected following variables: *AntibioticUsage, APGAR, AspirDurRat, BirthOption, BirthPlaceOGU, BirthWeek, ENDur, ENDurRat, Gender, HBM, HospDurLong2, ICUDurRat, IntubDurRat, LBW, LowAPGAR, Ncpap, NcpapDur, NgUseDur, PerCatDur, PerCatDurRat, TPNDur, TPNDurRat, UrinCatheter, Weight,* and *WeightBWeekRat.* In addition to the variables selected in holdout sampling, new variables related to treatment, patient and medical problem (*HBM,* indicator of hyperbilirubinemia) were selected by the models in cross validation.

Table 21 demonstrates the goodness of fit statistics provided by averaging the training and test folds. In comparison to holdout model, the difference between training and test sample results is lower. This indicates that the model handled the overfitting problem better than the first model built with holdout sampling.

Table 21. Entropy Cross Validation Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.13	0.15
Area under ROC	0.79	0.70
Gini Coefficient	0.57	0.39
Error Rate	0.17	0.18
Accuracy	0.83	0.82
Sensitivity	0.28	0.27
Specificity	0.99	0.97

The ROC curve of training sample represented in Figure 22 is lower than training sample ROC curve of holdout sample, indicating an adjustment in the model against overfitting.

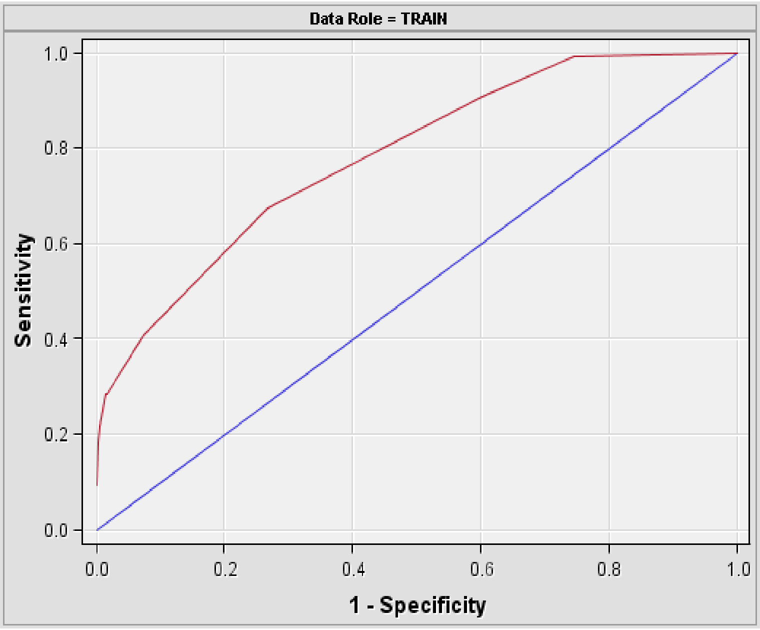

Figure 22. Entropy cross validation model ROC curve

Entropy Decision Tree Model with Bagging

As the base classifier, entropy decision tree with predefined parameters was applied on 10 training samples which were created with replacement. The results of bagged entropy decision tree are demonstrated in Table 22.

Table 22. Entropy Bagging Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.15
Area under ROC	0.72
Gini Coefficient	0.44
Error Rate	0.19
Accuracy	0.81
Sensitivity	0.23
Specificity	0.97

According to the results, bagged entropy tree shows considerable improvement in comparison to holdout model test sample. However, sensitivity measure is lower than the sensitivity of cross validation. Therefore, the bagging ROC curve is closer to the baseline, as illustrated in following Figure 23.

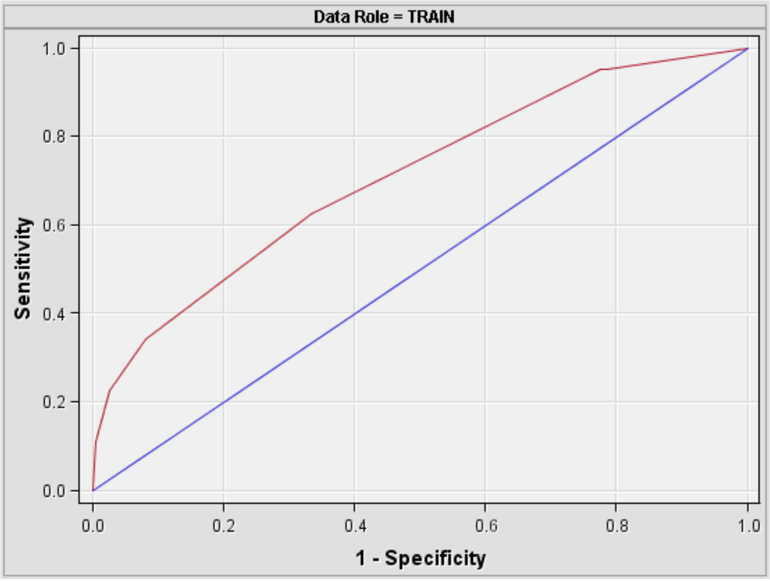

Figure 23. Entropy bagging model ROC curve

Entropy Decision Tree Model with Boosting

Boosting was applied on entropy decision tree with 10 iterations. The results of the method are represented in Table 23 and Figure 23 respectively. The boosted entropy tree has substantial improvement in sensitivity and deterioration in specificity. On the other hand, the error rate is relatively high.

Table 23. Entropy Boosting Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.21
Area under ROC	0.77
Gini Coefficient	0.53
Error Rate	0.34
Accuracy	0.66
Sensitivity	0.69
Specificity	0.64

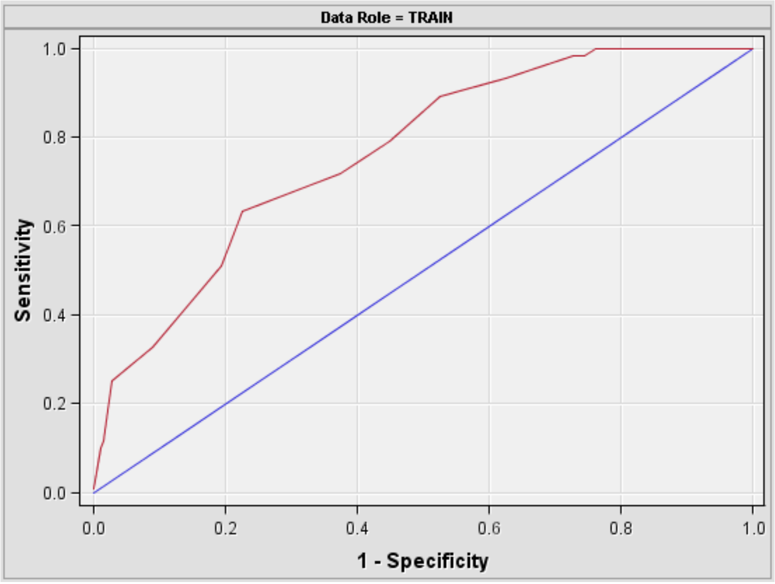

Figure 24. Entropy boosting model ROC curve

Because of the deviations in sensitivity and specificity measures, the ROC curve demonstrated in Figure 24 is fluctuating for boosted entropy decision trees.

Neural Networks

In the study, multilayer feed-forward perceptron with one hidden layer was built using backpropagation algorithm. The combination function was linear for both hidden and output layers. However, the activation functions were different at hidden and output layers. Hyperbolic tangent activation (transfer) function was used for hidden layers, where logistic activation function was used for output layer. As the dependent variable is binary, the error function was Bernoulli and the objective function was likelihood. The parameters of the model are given in following Table 24.

Table 24. Neural Networks Parameters

Parameter	Value
Number of Hidden Units	4
Maximum Iterations	100

As discussed in Chapter 4, variables that were selected by logistic regression model were used as input variables in order to prevent complex networks which poorly perform when trying to represent all the information in input variables. The input variables of neural networks are listed in Table 25.

Table 25. Input Variables of Neural Networks

Input Variables
ENDur
PerCatDur
UriCatDurRat
UrinCatheter

Continuous input variables *ENDur*, *PerCatDur* and *UriCatDurRat* were transformed as explained in Chapter 4. On the other hand, categorical variable *UrinCatheter* was recoded from {0,1} to {-1,1} in order to find a good local optimum.

The results of a neural network may depend on the initial values of the weights. Current backpropagation algorithm uses random initial weights. However, this may cause to find local minima. Therefore, a preliminary training was conducted using the selected input variables and parameters. The network was trained 5 times for 100 iterations in order to select the best estimates for the initial values of the weights. Consequently, these weights were used by the algorithm as the initial values of the weights for subsequent training.

Neural Networks Model with Holdout Sampling

Using predefined parameters and determined input variables, neural network model with holdout sampling was built. The generated weights for the selected iteration of the selected run are illustrated in a two dimensional diagram in Figure 25 below.

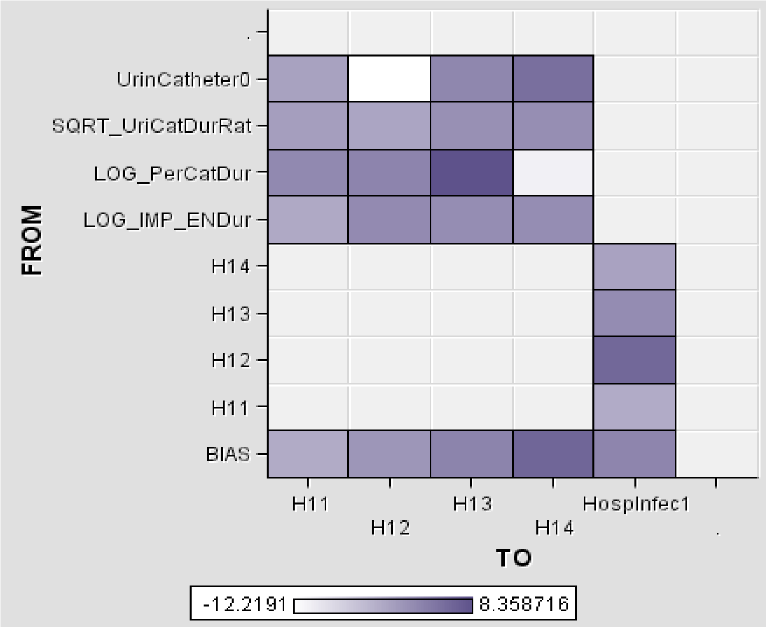

Figure 25. Final weights

The vertical axis is the source unit of the weight, where horizontal axis is the destination unit. The final weights are represented in Table 26 regarding Figure 24.

Table 26. Final Weights

EFFECT	FROM	TO	WEIGHT
LOG_IMP_ENDur -> H11	LOG_IMP_ENDur	H11	-19.357.594
LOG_PerCatDur -> H11	LOG_PerCatDur	H11	1.830.437
SQRT_UriCatDurRat -> H11	SQRT_UriCatDurRat	H11	-0.7137061
LOG_IMP_ENDur -> H12	LOG_IMP_ENDur	H12	16.456.205
LOG_PerCatDur -> H12	LOG_PerCatDur	H12	24.057.456
SQRT_UriCatDurRat -> H12	SQRT_UriCatDurRat	H12	-15.258.996
LOG_IMP_ENDur -> H13	LOG_IMP_ENDur	H13	13.178.359
LOG_PerCatDur -> H13	LOG_PerCatDur	H13	83.587.155
SQRT_UriCatDurRat -> H13	SQRT_UriCatDurRat	H13	10.005.458

Table 26. Continued

EFFECT	FROM	TO	WEIGHT
LOG_IMP_ENDur -> H14	LOG_IMP_ENDur	H14	12.848.623
LOG_PerCatDur -> H14	LOG_PerCatDur	H14	-1.046.493
SQRT_UriCatDurRat -> H14	SQRT_UriCatDurRat	H14	12.565.884
UrinCatheter0 -> H11	UrinCatheter0	H11	-11.231.683
UrinCatheter0 -> H12	UrinCatheter0	H12	-12.219.073
UrinCatheter0 -> H13	UrinCatheter0	H13	21.106.986
UrinCatheter0 -> H14	UrinCatheter0	H14	49.130.959
BIAS -> H11	BIAS	H11	-21.900.367
BIAS -> H12	BIAS	H12	0.3168653
BIAS -> H13	BIAS	H13	24.217.234
BIAS -> H14	BIAS	H14	60.117.869
H11 -> HospInfec1	H11	HospInfec1	-2.267.713
H12 -> HospInfec1	H12	HospInfec1	5.795.558
H13 -> HospInfec1	H13	HospInfec1	1.402.327
H14 -> HospInfec1	H14	HospInfec1	-12.014.169
BIAS -> HospInfec1	BIAS	HospInfec1	22.577.994

The goodness of fit statistics represented in Table 27 indicates 4% of decrease in the accuracy of the model when applied on the test sample. In addition, sensitivity, specificity, area under ROC and gini coefficient decreased considerably in the test sample.

Table 27. Neural Networks Holdout Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.12	0.16
Area under ROC	0.81	0.67
Gini Coefficient	0.61	0.35
Error Rate	0.15	0.19
Accuracy	0.85	0.81
Sensitivity	0.35	0.27
Specificity	0.99	0.97

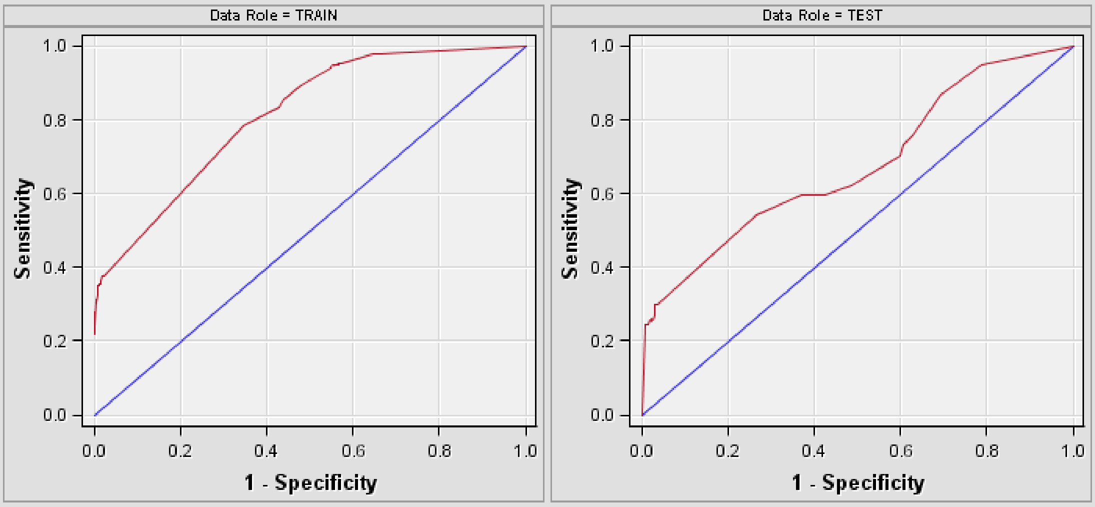

Figure 26. Neural networks holdout model ROC curve

The decrease in sensitivity and specificity measures of test sample drew down the ROC curve as illustrated in Figure 26.

Neural Networks Model with Cross Validation

The network was trained with 10-fold cross validation using the predefined parameters and input variables. The results showed that the model with cross validation is more consistent in terms of training and test sample statistics when compared to holdout model. The goodness of fit statistics and ROC curve are given in Table 28 and Figure 27 respectively.

Table 28. Neural Networks Cross Validation Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.13	0.14
Area under ROC	0.79	0.74
Gini Coefficient	0.59	0.48
Error Rate	0.17	0.17
Accuracy	0.83	0.83
Sensitivity	0.32	0.30
Specificity	0.98	0.97

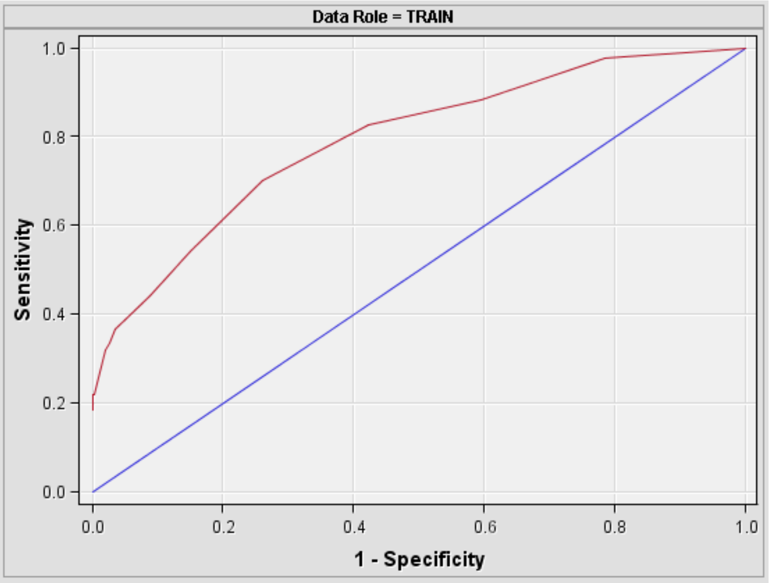

Figure 27. Neural networks cross validation model ROC curve

The ROC curve provided by averaging the training samples showed a slight improvement towards the upper left corner of the graph regarding the training and test samples' sensitivity and specificity measures.

Neural Networks Model with Bagging

A neural network with predefined parameters was applied on 10 training samples that were created with replacement. The results of bagged neural network are illustrated in Table 29.

Table 29. Neural Networks Bagging Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.13
Area under ROC	0.76
Gini Coefficient	0.51
Error Rate	0.17
Accuracy	0.83
Sensitivity	0.33
Specificity	0.98

The accuracy measure results of bagged networks are similar to those found by cross validation. However, according to area under ROC curve and gini coefficient it is seen that probabilities assigned to each observation are considerably different from cross validation. The ROC curve of bagged network is demonstrated in Figure 28.

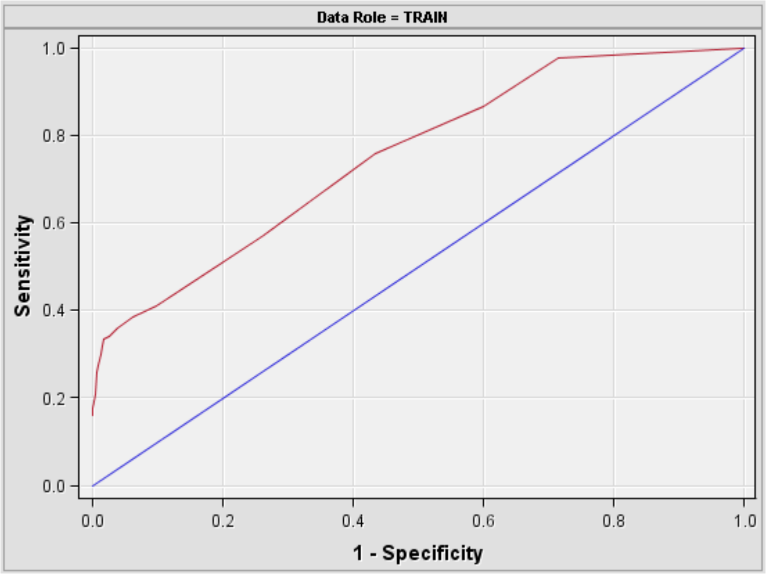

Figure 28. Neural networks bagging model ROC curve

Neural Networks Model with Boosting

Boosting was applied on neural networks with 20 iterations. The results of boosted networks are represented in Table 30 and Figure 29 respectively. A substantial improvement in sensitivity and a slight deterioration in specificity are observed. Although the area under ROC curve and gini coefficient decrease, the accuracy is quite high.

Table 30. Neural Networks Boosting Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.21
Area under ROC	0.70
Gini Coefficient	0.41
Error Rate	0.17
Accuracy	0.83
Sensitivity	0.44
Specificity	0.94

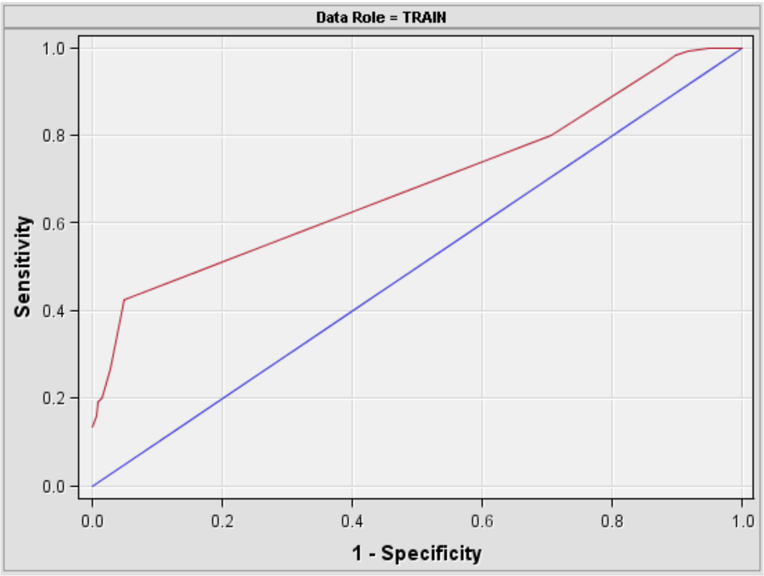

Figure 29. Neural networks boosting model ROC curve

Considering the classification measure sensitivity and specificity, the decrease in

ROC curve can be explained by the change in the distribution of assigned probabilities in comparison to holdout and cross validation method.

Logistic Regression

Logistic regression models are widely used for classification in medical data mining. In the study, stepwise model selection method was adopted in order to handle collinearity which indicates the relationship between two or more continuous independent variables. Collinearity may cause the coefficient estimates to change erratically in response to small changes in the model or the data. Correlation coefficients between the variables also help to identify the collinearity. Appendix D represents the Correlation Table among continuous variables.

Logistic regression was conducted with logit transformation function. No polynomial terms were used, only the main variables entered the stepwise selection method. There were 70 input variables for the logistic regression model which are listed in the following Table 31.

Table 31. Input Variables of Logistic Regression

Input Variables		
APGAR	Gender	PerCatDurRat
ARF	HBM	Phlebotomy
AgeDay	HospDur	PhototherDur
AntibioticUsage	HydropsFetalis	Polycythemia
AspirDur	ICUDur	Premature
AspirDurRat	ICUDurRat	Preterm
BirthHealthCenter	IUGR	RDS
BirthOption	ImmunoSupp	SteroidUsage
BirthPlaceBCent	IntubDur	SurgInterv

Table 31. Continued

Input Variables		
BirthPlaceHome	IntubDurRat	TPNDur
BirthPlaceHosp	Invasive	TPNDurRat
BirthPlaceOGU	LBW	TTN
BirthPlaceOth	LowAPGAR	Twin
BirthWeek	LowGISBleed	UmbCatDur
CatRelBloodFlow	MecAspSyn	UriCatDurRat
CentCatDur	MechVent	UriCatRelUrinary
ChestTube	NcpapDur	UrinCatheter
Chorioamnio	NgUseDur	VenDur
CongAno	OgUseDur	VenRelPneum
CutDown	Oligohyd	Weight
DetailBirthPlace	PAsphyxia	WeightBWeekRat
EMR	PPVDur	WrappedCord
ENDur	PerCatDur	h2blockPPI
ENDurRat		

In the study two logistic regression models were built. First model was built on the whole data set; second model was built with cross validation method. Bagging and boosting were not performed on logistic regression while both methods require unstable base classifiers such as decision trees and neural networks.

Logistic Regression Model

First logistic regression model was built on the whole data set after handling missing variables with decision tree induction as mentioned in Chapter 4. All the variables listed in Table 31 entered the model. A summary of stepwise selection is given in the following Table 32.

Table 32. Summary of Stepwise Selection

Summary of Stepwise Selection						
Step	Effect		Variables in model	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed				
1	PerCatDur		1	69.7127		<.0001
2	UrinCatheter		2	33.8498		<.0001
3	UriCatDurRat		3	10.0355		0.0015
4	IMP_ENDur		4	10.4483		0.0012
5	HydropsFetalis		5	7.0712		0.0078
6		HydropsFetalis	4		0.0113	0.9154

According to Table 32, model building terminated because the last effect entered was removed by the Wald test criterion where significance level 0.05 exceeded. The selected model was the model trained in Step 6 and it consists of the following variables: *ENDur*, *PerCatDur*, *UriCatDurRat*, and *UrinCatheter*.

Likewise, Figure 30 illustrates the average squared error in training set for each step in variable selection. The vertical blue line shows the step where the variable selection stopped.

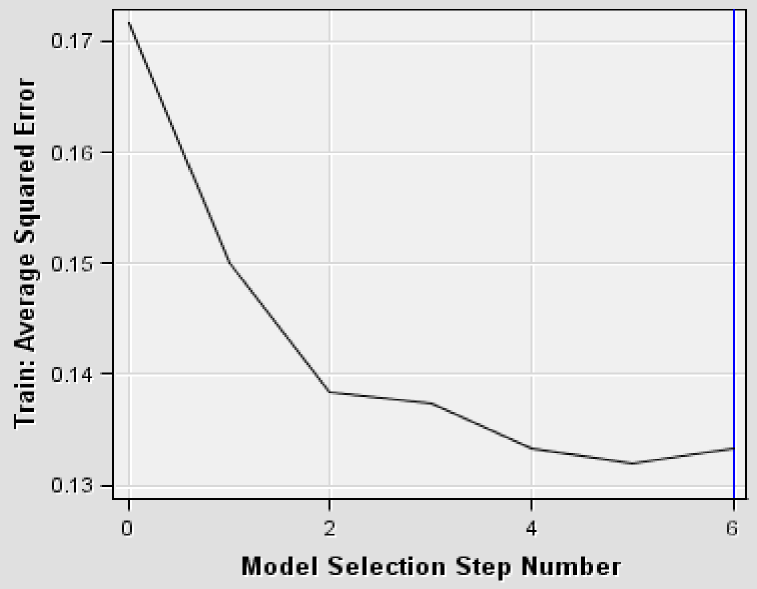

Figure 30. Average squared error in stepwise selection

The final coefficients and wald chi-square values of the selected variables are given in Table 33. According to the coefficients, it is seen that enteral nutrition duration (*ENDur*) and peripheral catheher duration (*PerCatDur*) are positively correlated with hospital infections. In addition, urinary catheter duration ratio (*UriCatDurRat*) and not to use urinary catheters (*UrinCatheter* = 0) are negatively correlated with the independent variable.

Table 33. Coefficients of the Variables

Variable	Coefficient	Wald Chi-Square	Pr > ChiSq
Intercept	0.6646	0.71	0.4003
IMP_ENDur	0.0694	10.21	0.0014
PerCatDur	0.0396	8.96	0.0028
UriCatDurRat	-7.047	6.69	0.0097
UrinCatheter (0)	-2.8067	12.63	0.0004

Following Table 34 provides -2LL values of the model with only intercept and with all selected variables to compare the models and the result of likelihood ratio chi-square test.

Table 34. Likelihood Ratio Test

-2 Log Likelihood		Likelihood Ratio Chi-Square	DF	Pr > ChiSq
Intercept Only	Intercept & Covariates			
574.583	467.19	107.3927	4	<.0001

Using -2LL values the pseudo R-Square was calculated as 0.19, which was considered significant. Likewise, the likelihood ratio chi-square showed that the model is significant for 0.05 significance value. The goodness of fit statistics for logistic regression model is given in the following Table 35.

Table 35. Logistic Regression Model Goodness of Fit Statistics

Statistics	Training
Average Squared Error	0.13
Area under ROC	0.76
Gini Coefficient	0.52
Error Rate	0.17
Accuracy	0.83
Sensitivity	0.31
Specificity	0.97

According to the goodness of fit statistics, the model has a sufficient accuracy. However, the sensitivity is low which indicates that the model poorly predicts the infected newborns. The following Figure 31 illustrates ROC curve of the model. It is seen that the curve has an increase at the beginning, however stayed parallel to baseline after it reaches 0.31 sensitivity.

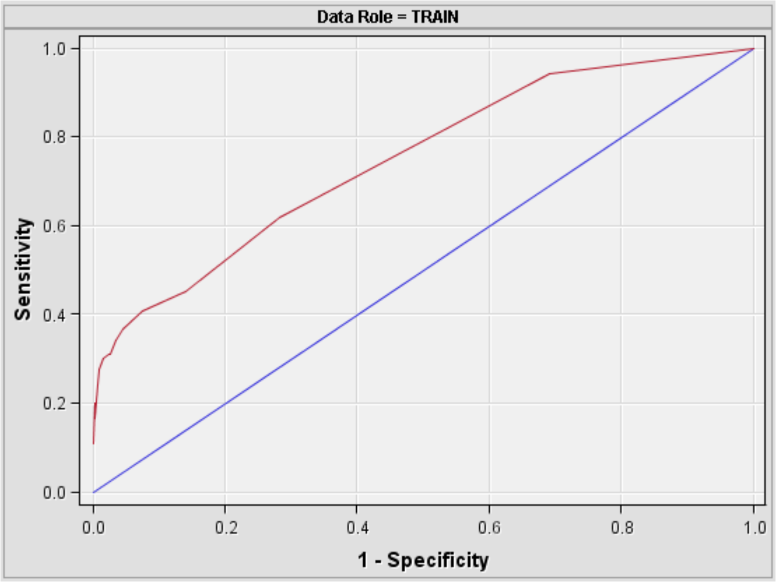

Figure 31. Logistic regression model ROC curve

Logistic Regression Model with Cross Validation

Logistic regression model with cross validation selected 10 variables related to treatment, patient and medical problem: *AntibioticUsage, BirthPlaceOGU, EMR, ENDurRat, PerCatDur, Polycythemia, TPNDurRat, UriCatDurRat, UrinCatheter,* and *WeightBWeekRat*. Table 36 demonstrates the goodness of fit statistics provided by averaging the training and test folds.

Table 36. Logistic Regression Cross Validation Model Goodness of Fit Statistics

Statistics	Training	Test
Average Squared Error	0.13	0.14
Area under ROC	0.77	0.75
Gini Coefficient	0.53	0.49
Error Rate	0.17	0.18
Accuracy	0.83	0.82
Sensitivity	0.30	0.32
Specificity	0.98	0.96

The ROC curve of the logistic regression with cross validation is given in Figure 32 below. In comparison to the first model, the curve shows an improvement in upper left direction.

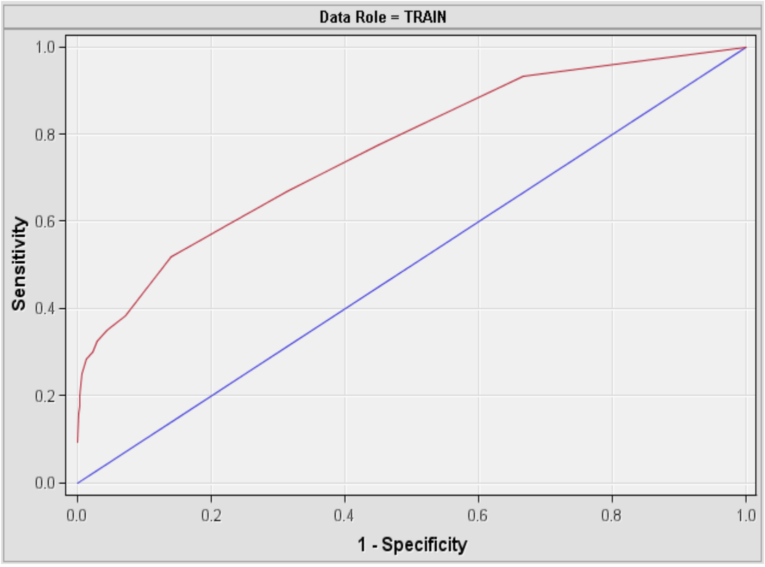

Figure 32. Logistic regression cross validation model ROC curve

Model Comparison

In random sampling, the error rate of training sample does not represent the true error rate of model's universe when modeling small data sets. Because of this, the generalization error is tried to be estimated with resampling methods. As the hospital infections data set was small, 10-fold cross validation method was selected to compare the performance of models in the study.

Variables selected by each cross validation model are listed in Table 37. As mentioned before in Chapter 4, neural network used variables that were selected by logistic regression model applied on whole data set.

The table shows that 39 different variables were selected from the whole variable list given in Table 1. Among these variables *AntibioticUsage, ENDur, ENDurRat, PerCatDur, TPNDur, TPNDurRat, UriCatDurRat, UrinCatheter* and *WeightBWeekRat* were chosen by more than two models. It is seen that CART and entropy models selected more than 20 variables. However, CART model selected 7 main variables to which an asterisk (*) is assigned in Table 37. The rest 22 variables are used as surrogate rules in case of missing values.

Table 37. Variables selected by Models

Variable	Model					Number of Models
	Chaid	Cart	Entropy	Neural Network	Logistic Regression	
AgeDay	□	✓	□	□	□	1
AntibioticUsage	✓	✓	✓	□	✓	4
APGAR	□	✓	✓	□	□	2
Aspiration	□	✓	□	□	□	1
AspirDur	□	✓	□	□	□	1
AspirDurRat	□	✓*	✓	□	□	2
BirthOption	□	□	✓	□	□	1
BirthPlaceOGU	□	□	✓	□	✓	2
BirthWeek	□	✓	✓	□	□	2
EMR	□	□	□	□	✓	1
ENDur	□	✓	✓	✓	□	3
ENDurRat	✓	✓*	✓	□	✓	4
EnteralNut	✓	✓	□	□	□	2
Gender	□	□	✓	□	□	1
HBM	□	□	✓	□	□	1
HospDur	□	✓	□	□	□	1
HospDurLong2	✓	□	✓	□	□	2
ICUDur	□	✓	□	□	□	1
ICUDurRat	□	□	✓	□	□	1
IntubDur	□	✓	□	□	□	1
IntubDurRat	□	✓	✓	□	□	2
LBW	□	✓	✓	□	□	2
LowAPGAR	✓	□	✓	□	□	2
Ncpap	□	□	✓	□	□	1
NcpapDur	□	✓	✓	□	□	2
NgUseDur	□	✓	✓	□	□	2
OgUseDur	□	✓	□	□	□	1
PerCatDur	✓	✓*	✓	✓	✓	5
PerCatDurRat	□	✓*	✓	□	□	2
PeripCatheter	□	✓	□	□	□	1
Polycythemia	□	□	□	□	✓	1
TPNDur	✓	✓*	✓	□	□	3
TPNDurRat	□	✓	✓	□	✓	3
UmbCatDur	□	✓	□	□	□	1
UriCatDur	□	✓*	□	□	□	1
UriCatDurRat	□	✓	□	✓	✓	3
UrinCatheter	✓	✓	✓	✓	✓	5
Weight	□	✓*	✓	□	□	2
WeightBWeekRat	□	✓	✓	□	✓	3
Number of Variables	8	29	25	4	10	

There are several criteria that can be used to compare the performance of classification models. As mentioned before in Chapter 3 while assessing the performance of medical data mining models, the error rate, accuracy and sensitivity measures have higher importance compared to others. In the study, specificity, area under ROC curve, gini coefficient and average squared error measures were also taken into consideration and the best model decision was given based on the test sample performance.

A good model is expected to have both low training and low test error. Moreover, overfitting can be determined via the difference between training and test error of two compared models. According to training sample, the model with lowest error rate, highest accuracy and sensitivity is CHAID decision tree. The second model is neural network and the third model is logistic regression. The training sample goodness of fit statistics and the ROC curve provided by training folds are given in Table 38 and Figure 33 respectively.

Table 38. Training Sample Goodness of Fit Statistics by Model

Statistics	CHAID	CART	Entropy	Neural Network	Logistic Regression
Average Squared Error	0.14	0.14	0.13	0.13	0.13
Area under ROC	0.73	0.66	0.79	0.79	0.77
Gini Coefficient	0.45	0.32	0.57	0.59	0.53
Error Rate	0.17	0.17	0.17	0.17	0.17
Accuracy	0.83	0.83	0.83	0.83	0.83
Sensitivity	0.33	0.28	0.28	0.32	0.30
Specificity	0.97	0.99	0.99	0.98	0.98

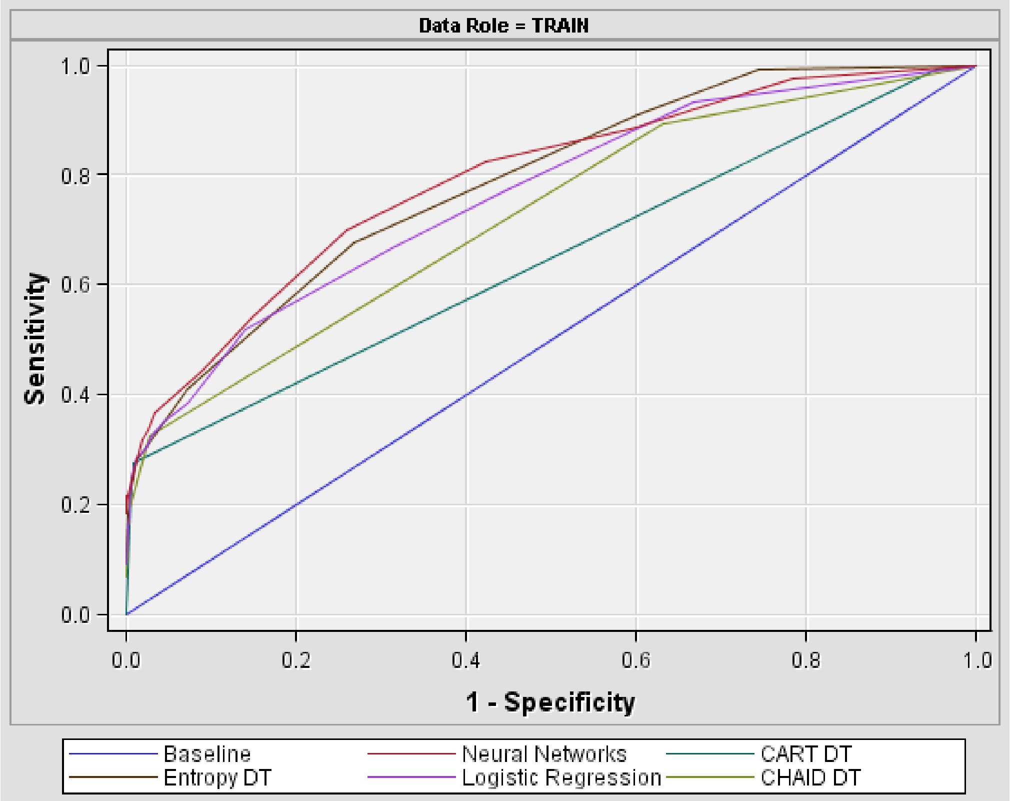

Figure 33. Model comparison cross validation ROC curves

Beyond test samples, there are two successful models according to determined criteria. These are CHAID and neural network models. CHAID model is superior to neural network in terms of specificity by only 1%. However, neural network model is superior to CHAID model in terms of area under ROC curve and gini coefficient. Test sample results are illustrated in following Figure 39.

Table 39. Test Sample Goodness of Fit Statistics by Model

Statistics	CHAID	CART	Entropy	Neural Network	Logistic Regression
Average Squared Error	0.14	0.15	0.15	0.14	0.14
Area under ROC	0.68	0.65	0.7	0.74	0.75
Gini Coefficient	0.36	0.31	0.39	0.48	0.49
Error Rate	0.17	0.19	0.18	0.17	0.18
Accuracy	0.83	0.81	0.82	0.83	0.82
Sensitivity	0.30	0.26	0.27	0.30	0.32
Specificity	0.98	0.97	0.97	0.97	0.96

According to these criteria, neural network model is found to be the most accurate model. Furthermore, probability distribution of models demonstrated in Figure 37 in Appendix E shows that neural network model distinguishes infected and noninfected newborns better due to the distance between cumulative percentage curves. Although, it is not to be omitted that for medical data mining, accessible interpretation that explains input-output relations is of extreme importance. Therefore, CHAID is a good candidate model to be applied in terms of explicit rules, short construction and running time, robustness to missing data, and ability to incorporate pre-existing knowledge.

CHAPTER 6

CONCLUSION

The hospital infections are important health care problems. They cause high morbidity, mortality and economic burden. Newborns are more exposed to these infections as their floras which protect them from disease-producing bacteria are not formed as they are born. The aim of this study is to discover a pattern of hospital infections in newborn ICUs with the data collected by Department of Clinical Microbiology and Infectious Diseases, Eskişehir Osmangazi University, Faculty of Medicine.

The methodology followed in this study is KDD process. In data cleaning and preprocessing step, inconsistencies caused by manual data entry and missing values are handled. In data reduction and transformation step, for different classification techniques, different input variables are selected and different data transformation techniques are applied.

In modeling step, the aim is to find the most accurate model that classifies newborns successfully. CHAID, CART and entropy decision trees, neural network and logistic regression techniques are applied. Because of the small sample size, in addition to holdout sampling, models are built and compared with cross validation. The best model is determined by the lowest error rate, highest accuracy and sensitivity in test set. Besides, area under ROC curve, gini coefficient, average squared error, and specificity measures are taken into consideration.

Neural network and CHAID decision tree models present considerable classification performance. Both models have the same accuracy and sensitivity on the test set. However, neural network model is superior to CHAID in terms of area under ROC curve and gini coefficient. Still, CHAID model is a good candidate as accessible interpretation that explains input-output relations is very important in medical applications. Moreover, CHAID model provides short construction and running time, robustness to missing data and ability to incorporate pre-existing knowledge.

In addition to holdout sampling and cross validation, accuracy increasing methods bagging and boosting are applied to decision trees and neural networks. Bagged CHAID and boosted neural network models reached the highest sensitivity rates with considerable accuracy.

Variables *AntibioticUsage, ENDurRat, EnteralNut, HospDurLong2, LowAPGAR, PerCatDur, TPNDur,* and *UrinCatheter* are selected by CHAID models. Among these variables *AntibioticUsage, ENDurRat, PerCatDur, TPNDur* and *UrinCatheter* are selected by more than two models in the study. Variables indicate that during the treatment of the main disease, applied methods such as catheters may increase the risk of hospital infections. Moreover, patient characteristics such as low APGAR score and lengthened hospital stay may trigger the hospital infections.

In summary, the surveillance of hospital infections in newborn ICUs has an extreme importance to prevent the outbreaks. In addition to infection control programs which prevent hospital infections, systems that can also expose infection signals are very important. Detecting risky or infected newborns accurately will help to reduce morbidity and mortality with on time and right treatments and likewise, determining less risky - noninfected newborns will reduce the economic burden of hospital infections.

In further studies, more observations regarding hospital infections in newborn ICUs should be analyzed with classification algorithms. Moreover, the sample should be collected from several hospitals to be able to eliminate the local effects. Thus for different hospital infection types, different models can be built. In addition, patient-care related risk factors such as disinfection and antisepsis usage frequency and if possible, external factors such as season, air-conditioning etc. should be taken into consideration for hospital infection detection in newborn ICUs.

APPENDICES

A. MEDICAL DEFINITIONS

Nosocomial: Originating or taking place in a hospital, especially in reference to an infection (Medicalterms.com, 2008).

Incubation period: In medicine, the time from the moment of exposure to an infectious agent until signs and symptoms of the disease appear (Medterms.com, 2008).

CDC National Nosocomial Infection Surveillance (NNIS): System in the United States has provided standardized methods for collecting and comparing health care-associated infection rates and the national benchmark infection rate data for inter- and intrahospital comparisons since the 1970s. (Pittet, 2005)

Surveillance: Surveillance is defined as ‘‘the ongoing, systematic collection, analysis, and interpretation of data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know.’’(Pittet, 2005)

Colonization: Microorganism is present in host and increases its population. Colonization is different from infection in terms of not causing illness (tipterimleri.com, 2008).

Neonate: Newborn baby, the first 28 days of a baby.

Flora: Microorganisms that live on or within a body to compete with disease-producing microorganisms and provide a natural immunity against certain infections (medical-dictionary.thefreedictionary.com, 2010).

Apgar Score: Results of a method to assess the health of a newborn according to five main criteria (Appearance, Pulse, Grimace, Activity, and Respiration) immediately after birth. Each criteria is scored between 0 and 2 and sum of these scores are evaluated as Apgar score. A normal child gets a score between 7-10. Apgar score is observed twice: one minute after birth and five minutes after birth.

Celebral Palsy: A non-progressive abnormality of motor function that can cause seizures, subnormality, blindness, deafness, nutrition or swallow disorders (tipterimleri.com, 2008).

Acidosis: Blood PH is abnormally low because of acid accumulation or alkaline reserves' deplation in blood or tissues.

Gestational Age: Age of embryo in weeks.

Spa typing: DNA sequence analysis of the protein A gene variable repeat region (spa typing) provides a rapid and accurate method to discriminate Staphylococcus aureus outbreak (medical-dictionary.thefreedictionary.com, 2010).

Invasive procedure: A medical procedure which penetrates or breaks the skin or a body cavity, i.e., it requires a perforation, an incision, a catheterization, etc. into the body (medterms.com, 2008).

Sepsis, Septicemia: The presence of bacteria or other infectious organisms or their toxins in the blood (septicemia) or in other tissue of the body. It is commonly called "blood stream infection". Babies with sepsis may be listless, overly sleepy, floppy, weak, and very pale.

Early membrane rupture (EMR): Early detachment of the placenta from the uterus (medical-dictionary.thefreedictionary.com, 2010).

Perinatal Asphyxia: Respiratory failure in the newborn, a condition caused by the inadequate intake of oxygen before, during, or just after birth (medical-dictionary.thefreedictionary.com, 2010).

Hydrops Fetalis: A condition in which a fetus or newborn baby accumulates fluids, causing swollen arms and legs and impaired breathing (medical-dictionary.thefreedictionary.com, 2010).

Polycythemia: An increase in red cell mass caused by increased erythropoiesis, over 65% (medical-dictionary.thefreedictionary.com, 2010).

Acute Renal Failure: An abrupt decline in renal function (medical-dictionary.thefreedictionary.com, 2010).

Meconium Aspiration Syndrome: The respiratory complications resulting from the passage and aspiration of meconium prior to or during delivery (medical-dictionary.thefreedictionary.com, 2010).

Chorioamnionitis: Inflammation of the amniotic membranes caused by infection (medical-dictionary.thefreedictionary.com, 2010).

Oligohydramnios: An abnormally small amount or absence of amniotic fluid (medical-dictionary.thefreedictionary.com, 2010).

Congenital Anomaly: A developmental anomaly present at birth (medical-dictionary.thefreedictionary.com, 2010).

Immunosuppression: Suppression of the immune response, as by drugs or radiation, in order to prevent the rejection of grafts or transplants or control autoimmune diseases (medical-dictionary.thefreedictionary.com, 2010).

Urinary Catheter: The insertion of a catheter into a patient's bladder (medical-dictionary.thefreedictionary.com, 2010).

Peripheral Catheter: An intravenous line placed in a vein (oncolink.org, 2010).

Umbilical Catheter: A procedure in which a radiopaque catheter is passed through an umbilical artery to provide a newborn with parenteral fluid, to obtain blood samples, or both, or through the umbilical vein for an exchange transfusion or the emergency administration of drugs, fluids, or volume expanders (medical-dictionary.thefreedictionary.com, 2010).

Total Parenteral Feeding: A way of supplying all the nutritional needs of the body by bypassing the digestive system and dripping nutrient solution directly into a vein (medical-dictionary.thefreedictionary.com, 2010).

Enteral Nutrition: the delivery of nutrients in liquid form directly into the stomach, duodenum, or jejunum (medical-dictionary.thefreedictionary.com, 2010).

Nasogastric Intubation: The placement of a nasogastric tube through the nose into the stomach (medical-dictionary.thefreedictionary.com, 2010).

Orogastric Intubation: passing a stomach tube via the mouth (medical-dictionary.thefreedictionary.com, 2010).

H2 Blockers or Proton-Pump Inhibitors (PPIs): Act by stopping the pathway that leads to the secretion of stomach acid / drugs reduce the secretion of gastric (stomach) acid (medical-dictionary.thefreedictionary.com, 2010).

Steroid Usage: A natural body substance that often is given to women before delivering a very premature infant to stimulate the fetal lungs to produce surfactant, hopefully preventing RDS (medical-dictionary.thefreedictionary.com, 2010).

Aspiration: Removal by suction (medical-dictionary.thefreedictionary.com, 2010).

Intubation: The insertion of a tube into a body canal or hollow organ (medical-dictionary.thefreedictionary.com, 2010).

Mechanical Ventilation: Breathing that accomplished by extrinsic means (medical-dictionary.thefreedictionary.com, 2010).

Nasal Continuous Positive Airway Pressure (nCPAP): A ventilation device that blows a gentle stream of air into the nose to keep the airway open (medical-dictionary.thefreedictionary.com, 2010).

Tracheostomy: Surgical construction of a respiratory opening in the trachea (medical-dictionary.thefreedictionary.com, 2010).

Cut down: Creation of a small incised opening, especially over a vein (medical-dictionary.thefreedictionary.com, 2010).

Chest tube: A catheter inserted through the rib space of the thorax into the pleural space (medical-dictionary.thefreedictionary.com, 2010).

Positive pressure ventilation: Mechanical ventilation in which air is delivered into the airways and lungs under positive pressure (medical-dictionary.thefreedictionary.com, 2010).

Respiratory distress syndrome: An acute lung disease present at birth (medical-dictionary.thefreedictionary.com, 2010).

Exchange transfusion (blood exchange): Repetitive withdrawal of small amounts of blood and replacement with donor blood (medical-dictionary.thefreedictionary.com, 2010).

Phlebotomy: The act of drawing or removing blood from the circulatory system through a cut (medical-dictionary.thefreedictionary.com, 2010).

Phototherapy: A treatment for hyperbilirubinemia and jaundice in the newborn that involves the exposure of an infant's bare skin to intense fluorescent light (medical-dictionary.thefreedictionary.com, 2010).

B. CROSS TABULATIONS OF CATEGORICAL VARIABLES

In order to examine the relationship between independent categorical variables and dependent variable (HospInfec), cross tabulations with cell percentages were drawn.

Table 40. Cross Tabulations

Variables	Value	Type	HospInfec		Total
			0	1	
ARF	0	N	413	117	530
		PctN	75.78	21.47	97.25
	1	N	12	3	15
		PctN	2.2	0.55	2.75
	Total	N	425	120	545
		PctN	77.98	22.02	100
Aspiration	0	N	383	89	472
		PctN	70.27	16.33	86.61
	1	N	42	31	73
		PctN	7.71	5.69	13.39
	Total	N	425	120	545
		PctN	77.98	22.02	100
BirthHealthCenter	0	N	8	1	9
		PctN	1.47	0.18	1.65
	1	N	417	119	536
		PctN	76.51	21.83	98.34
	Total	N	425	120	545
		PctN	77.98	22.02	100
BirthOption	0	N	198	55	253
		PctN	36.33	10.09	46.42
	1	N	227	65	292
		PctN	41.65	11.93	53.58
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
BirthPlaceBCent	0	N	404	119	523
		PctN	74.13	21.83	95.96
	1	N	21	1	22
		PctN	3.85	0.18	4.03
	Total	N	425	120	545
		PctN	77.98	22.01	100
BirthPlaceHome	0	N	418	120	538
		PctN	76.7	22.02	98.72
	1	N	7	.	7
		PctN	1.28	.	1.28
	Total	N	425	120	545
		PctN	77.98	22.02	100
BirthPlaceHosp	0	N	224	73	297
		PctN	41.1	13.39	54.49
	1	N	201	47	248
		PctN	36.88	8.62	45.5
	Total	N	425	120	545
		PctN	77.98	22.02	100
BirthPlaceOGU	0	N	230	49	279
		PctN	42.2	8.99	51.19
	1	N	195	71	266
		PctN	35.78	13.03	48.81
	Total	N	425	120	545
		PctN	77.98	22.02	100
BirthPlaceOth	0	N	424	119	543
		PctN	77.8	21.83	99.63
	1	N	1	1	2
		PctN	0.18	0.18	0.36
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
BirthWGroup2	0	N	41	26	67
		PctN	7.52	4.77	12.29
	1	N	384	94	478
		PctN	70.46	17.25	87.71
	Total	N	425	120	545
		PctN	77.98	22.02	100
BirthWGroup4	1	N	4	6	10
		PctN	0.73	1.1	1.83
	2	N	37	20	57
		PctN	6.79	3.67	10.46
	3	N	97	29	126
		PctN	17.8	5.32	23.12
	4	N	287	65	352
		PctN	52.66	11.93	64.59
	Total	N	425	120	545
		PctN	77.98	22.02	100
CatRelBloodFlow	0	N	425	119	544
		PctN	77.98	21.83	99.81
	1	N	.	1	1
		PctN	.	0.18	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
CentCatheter	0	N	424	120	544
		PctN	77.8	22.02	99.82
	1	N	1	.	1
		PctN	0.18	.	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
ChestTube	0	N	418	117	535
		PctN	76.7	21.47	98.17
	1	N	7	3	10
		PctN	1.28	0.55	1.83
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
Chorioamnio	0	N	423	118	541
		PctN	77.61	21.65	99.26
	1	N	2	2	4
		PctN	0.37	0.37	0.74
	Total	N	425	120	545
		PctN	77.98	22.02	100
CongAno	0	N	365	103	468
		PctN	66.97	18.9	85.87
	1	N	60	17	77
		PctN	11.01	3.12	14.13
	Total	N	425	120	545
		PctN	77.98	22.02	100
CutDown	0	N	424	120	544
		PctN	77.8	22.02	99.82
	1	N	1	.	1
		PctN	0.18	.	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
DetailBirthPlace	1	N	195	71	266
		PctN	35.78	13.03	48.81
	2	N	201	47	248
		PctN	36.88	8.62	45.5
	3	N	21	1	22
		PctN	3.85	0.18	4.04
	4	N	7	.	7
		PctN	1.28	.	1.28
	5	N	1	1	2
		PctN	0.18	0.18	0.36
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
EMR	0	N	393	113	506
		PctN	72.11	20.73	92.84
	1	N	32	7	39
		PctN	5.87	1.28	7.16
	Total	N	425	120	545
		PctN	77.98	22.01	100
EnteralNut	.	N	27	13	40
		PctN	4.95	2.39	7.34
	0	N	242	41	283
		PctN	44.4	7.52	51.93
	1	N	156	66	222
		PctN	28.62	12.11	40.73
	Total	N	425	120	545
		PctN	77.98	22.02	100
Gender	0	N	215	55	270
		PctN	39.45	10.09	49.54
	1	N	210	65	275
		PctN	38.53	11.93	50.46
	Total	N	425	120	545
		PctN	77.98	22.02	100
h2blockPPI	0	N	410	107	517
		PctN	75.23	19.63	94.86
	1	N	15	13	28
		PctN	2.75	2.39	5.14
	Total	N	425	120	545
		PctN	77.98	22.02	100
HBM	0	N	156	36	192
		PctN	28.62	6.61	35.23
	1	N	269	84	353
		PctN	49.36	15.41	64.77
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
HospDurLong2	0	N	257	42	299
		PctN	47.16	7.71	54.86
	1	N	168	78	246
		PctN	30.83	14.31	45.14
	Total	N	425	120	545
		PctN	77.98	22.02	100
HydropsFetalis	0	N	425	119	544
		PctN	77.98	21.83	99.81
	1	N	.	1	1
		PctN	.	0.18	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
ICU	0	N	381	95	476
		PctN	69.91	17.43	87.34
	1	N	44	25	69
		PctN	8.07	4.59	12.66
	Total	N	425	120	545
		PctN	77.98	22.02	100
ImmunoSupp	0	N	424	120	544
		PctN	77.8	22.02	99.82
	1	N	1	.	1
		PctN	0.18	.	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
Intubation	0	N	357	78	435
		PctN	65.5	14.31	79.81
	1	N	68	42	110
		PctN	12.48	7.71	20.18
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
Invasive	0	N	384	105	489
		PctN	70.46	19.27	89.73
	1	N	41	15	56
		PctN	7.52	2.75	10.28
	Total	N	425	120	545
		PctN	77.98	22.02	100
IUGR	0	N	402	114	516
		PctN	73.76	20.92	94.68
	1	N	23	6	29
		PctN	4.22	1.1	5.32
	Total	N	425	120	545
		PctN	77.98	22.02	100
LBW	1	N	66	16	82
		PctN	12.11	2.94	15.05
	2	N	336	97	433
		PctN	61.65	17.8	79.45
	3	N	23	7	30
		PctN	4.22	1.28	5.5
	Total	N	425	120	545
		PctN	77.98	22.02	100
LowAPGAR	.	N	75	31	106
		PctN	13.76	5.69	19.45
	0	N	337	85	422
		PctN	61.83	15.6	77.43
	1	N	13	4	17
		PctN	2.39	0.73	3.12
	Total	N	425	120	545
		PctN	77.98	22.02	100
LowGISBleed	0	N	424	119	543
		PctN	77.8	21.83	99.63
	1	N	1	1	2
		PctN	0.18	0.18	0.36
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
MecAspSyn	0	N	393	111	504
		PctN	72.11	20.37	92.48
	1	N	32	9	41
		PctN	5.87	1.65	7.52
	Total	N	425	120	545
		PctN	77.98	22.02	100
MechVent	0	N	356	78	434
		PctN	65.32	14.31	79.63
	1	N	69	42	111
		PctN	12.66	7.71	20.37
	Total	N	425	120	545
		PctN	77.98	22.02	100
NasogTube	0	N	390	94	484
		PctN	71.56	17.25	88.81
	1	N	35	26	61
		PctN	6.42	4.77	11.19
	Total	N	425	120	545
		PctN	77.98	22.02	100
Ncpap	0	N	316	73	389
		PctN	57.98	13.39	71.37
	1	N	109	47	156
		PctN	20	8.62	28.62
	Total	N	425	120	545
		PctN	77.98	22.02	100
Oligohyd	0	N	418	117	535
		PctN	76.7	21.47	98.17
	1	N	7	3	10
		PctN	1.28	0.55	1.83
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
Ncpap	0	N	316	73	389
		PctN	57.98	13.39	71.37
	1	N	109	47	156
		PctN	20	8.62	28.62
	Total	N	425	120	545
		PctN	77.98	22.02	100
Oligohyd	0	N	418	117	535
		PctN	76.7	21.47	98.17
	1	N	7	3	10
		PctN	1.28	0.55	1.83
	Total	N	425	120	545
		PctN	77.98	22.02	100
OrogTube	0	N	387	93	480
		PctN	71.01	17.06	88.07
	1	N	38	27	65
		PctN	6.97	4.95	11.93
	Total	N	425	120	545
		PctN	77.98	22.02	100
Pasphyxia	0	N	383	99	482
		PctN	70.28	18.17	88.45
	1	N	42	21	63
		PctN	7.71	3.85	11.56
	Total	N	425	120	545
		PctN	77.98	22.02	100
PeripCatheter	0	N	90	5	95
		PctN	16.51	0.92	17.43
	1	N	335	115	450
		PctN	61.47	21.1	82.57
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
OrogTube	0	N	387	93	480
		PctN	71.01	17.06	88.07
	1	N	38	27	65
		PctN	6.97	4.95	11.93
	Total	N	425	120	545
		PctN	77.98	22.02	100
Pasphyxia	0	N	383	99	482
		PctN	70.28	18.17	88.45
	1	N	42	21	63
		PctN	7.71	3.85	11.56
	Total	N	425	120	545
		PctN	77.98	22.02	100
PeripCatheter	0	N	90	5	95
		PctN	16.51	0.92	17.43
	1	N	335	115	450
		PctN	61.47	21.1	82.57
	Total	N	425	120	545
		PctN	77.98	22.02	100
Phlebotomy	0	N	424	120	544
		PctN	77.8	22.02	99.82
	1	N	1	.	1
		PctN	0.18	.	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
Phototherapy	0	N	155	40	195
		PctN	28.44	7.34	35.78
	1	N	270	80	350
		PctN	49.54	14.68	64.22
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
Polycythemia	0	N	414	114	528
		PctN	75.96	20.92	96.88
	1	N	11	6	17
		PctN	2.02	1.1	3.12
	Total	N	425	120	545
		PctN	77.98	22.02	100
PPV	0	N	390	105	495
		PctN	71.56	19.27	90.83
	1	N	35	15	50
		PctN	6.42	2.75	9.17
	Total	N	425	120	545
		PctN	77.98	22.02	100
Premature	0	N	282	66	348
		PctN	51.74	12.11	63.85
	1	N	143	54	197
		PctN	26.24	9.91	36.15
	Total	N	425	120	545
		PctN	77.98	22.02	100
Preterm	0	N	380	100	480
		PctN	69.72	18.35	88.07
	1	N	45	20	65
		PctN	8.26	3.67	11.93
	Total	N	425	120	545
		PctN	77.98	22.02	100
RDS	0	N	399	103	502
		PctN	73.21	18.9	92.11
	1	N	26	17	43
		PctN	4.77	3.12	7.89
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
SteroidUsage	0	N	410	105	515
		PctN	75.23	19.27	94.5
	1	N	15	15	30
		PctN	2.75	2.75	5.5
	Total	N	425	120	545
		PctN	77.98	22.02	100
SurgInterv	0	N	418	109	527
		PctN	76.7	20	96.7
	1	N	7	11	18
		PctN	1.28	2.02	3.3
	Total	N	425	120	545
		PctN	77.98	22.02	100
TPN	0	N	308	68	376
		PctN	56.51	12.48	68.99
	1	N	117	52	169
		PctN	21.47	9.54	31.01
	Total	N	425	120	545
		PctN	77.98	22.02	100
TPNLipid	0	N	307	68	375
		PctN	56.33	12.48	68.81
	1	N	118	52	170
		PctN	21.65	9.54	31.19
	Total	N	425	120	545
		PctN	77.98	22.02	100
Trakeos	0	N	425	120	545
		PctN	77.98	22.02	100
	Total	N	425	120	545
		PctN	77.98	22.02	100
TTN	0	N	400	115	515
		PctN	73.39	21.1	94.49
	1	N	25	5	30
		PctN	4.59	0.92	5.51
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
Twin	0	N	377	108	485
		PctN	69.17	19.82	88.99
	1	N	48	12	60
		PctN	8.81	2.2	11.01
	Total	N	425	120	545
		PctN	77.98	22.02	100
UmbCatheter	0	N	392	111	503
		PctN	71.93	20.37	92.3
	1	N	33	9	42
		PctN	6.06	1.65	7.71
	Total	N	425	120	545
		PctN	77.98	22.02	100
UriCatRelUrinary	0	N	425	119	544
		PctN	77.98	21.83	99.81
	1	N	.	1	1
		PctN	.	0.18	0.18
	Total	N	425	120	545
		PctN	77.98	22.02	100
UrinCatheter	0	N	423	106	529
		PctN	77.61	19.45	97.06
	1	N	2	14	16
		PctN	0.37	2.57	2.94
	Total	N	425	120	545
		PctN	77.98	22.02	100
VenRelPneum	.	N	1	.	1
		PctN	0.18	.	0.18
	0	N	424	116	540
		PctN	77.8	21.28	99.08
	1	N	.	4	4
		PctN	.	0.73	0.73
	Total	N	425	120	545
		PctN	77.98	22.02	100

Table 40. continued.

Variables	Value	Type	HospInfec		Total
			0	1	
WeightGroup22	0	N	275	67	342
		PctN	50.46	12.29	62.75
	1	N	150	53	203
		PctN	27.52	9.72	37.24
	Total	N	425	120	545
		PctN	77.98	22.02	100
WeightGroup3	0	N	27	16	43
		PctN	4.95	2.94	7.89
	1	N	123	37	160
		PctN	22.57	6.79	29.36
	2	N	275	67	342
		PctN	50.46	12.29	62.75
	Total	N	425	120	545
		PctN	77.98	22.02	100
WeightGroup5	1	N	5	2	7
		PctN	0.92	0.37	1.29
	2	N	22	14	36
		PctN	4.04	2.57	6.61
	3	N	62	18	80
		PctN	11.38	3.3	14.68
	4	N	61	19	80
		PctN	11.19	3.49	14.68
	5	N	275	67	342
		PctN	50.46	12.29	62.75
	Total	N	425	120	545
		PctN	77.98	22.02	100
WrappedCord	0	N	422	118	540
		PctN	77.43	21.65	99.08
	1	N	3	2	5
		PctN	0.55	0.37	0.92
	Total	N	425	120	545
		PctN	77.98	22.02	100

C. DECISION TREE STRUCTURES

Classification trees of holdout decision tree models are represented in following figures. Low (white) to high (dark blue) probabilities of correct classification are displayed using a graduated color ramp.

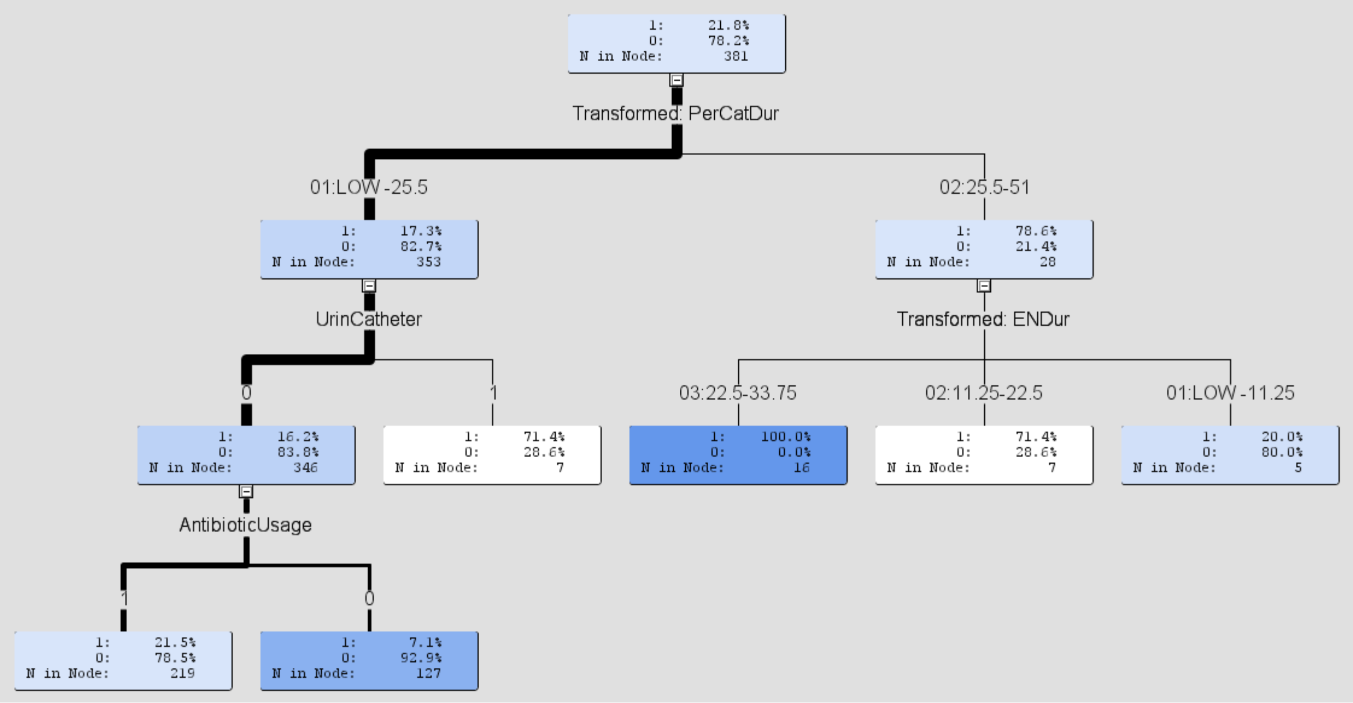

Figure 34. CHAID decision tree with holdout sampling

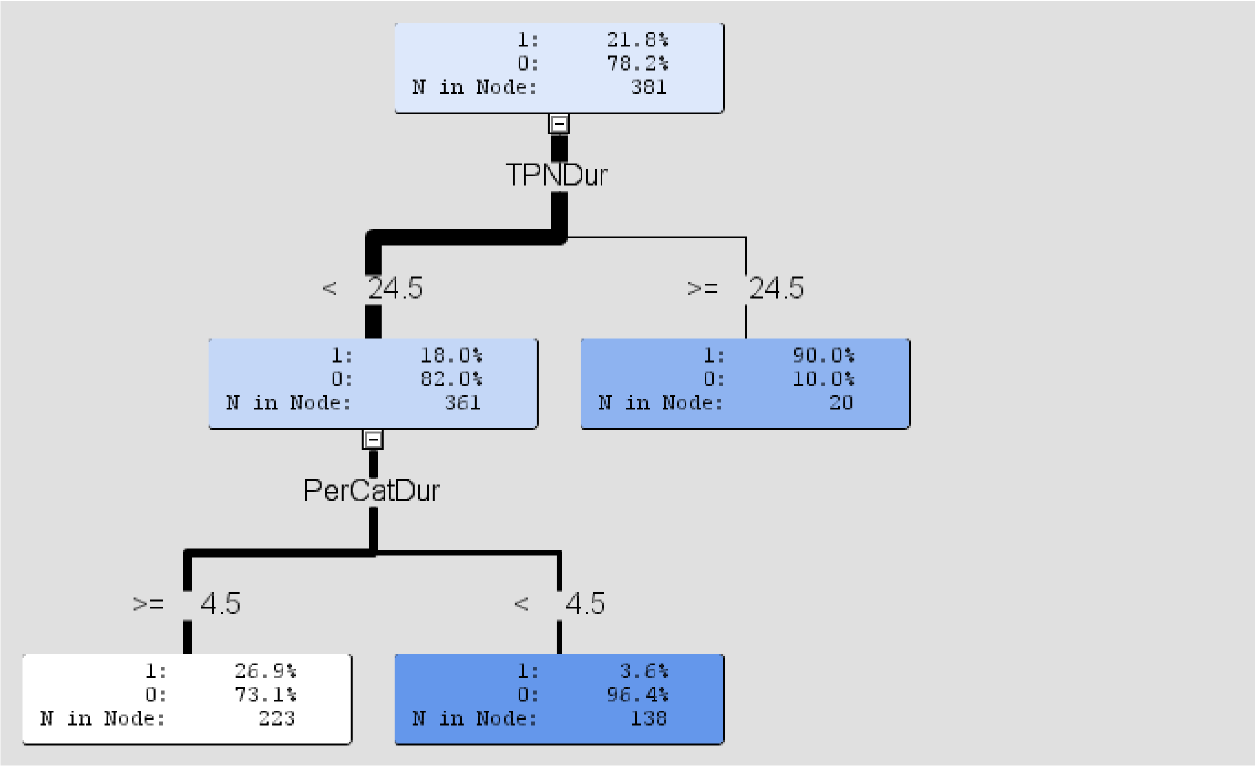

Figure 35. CART decision tree with holdout sampling

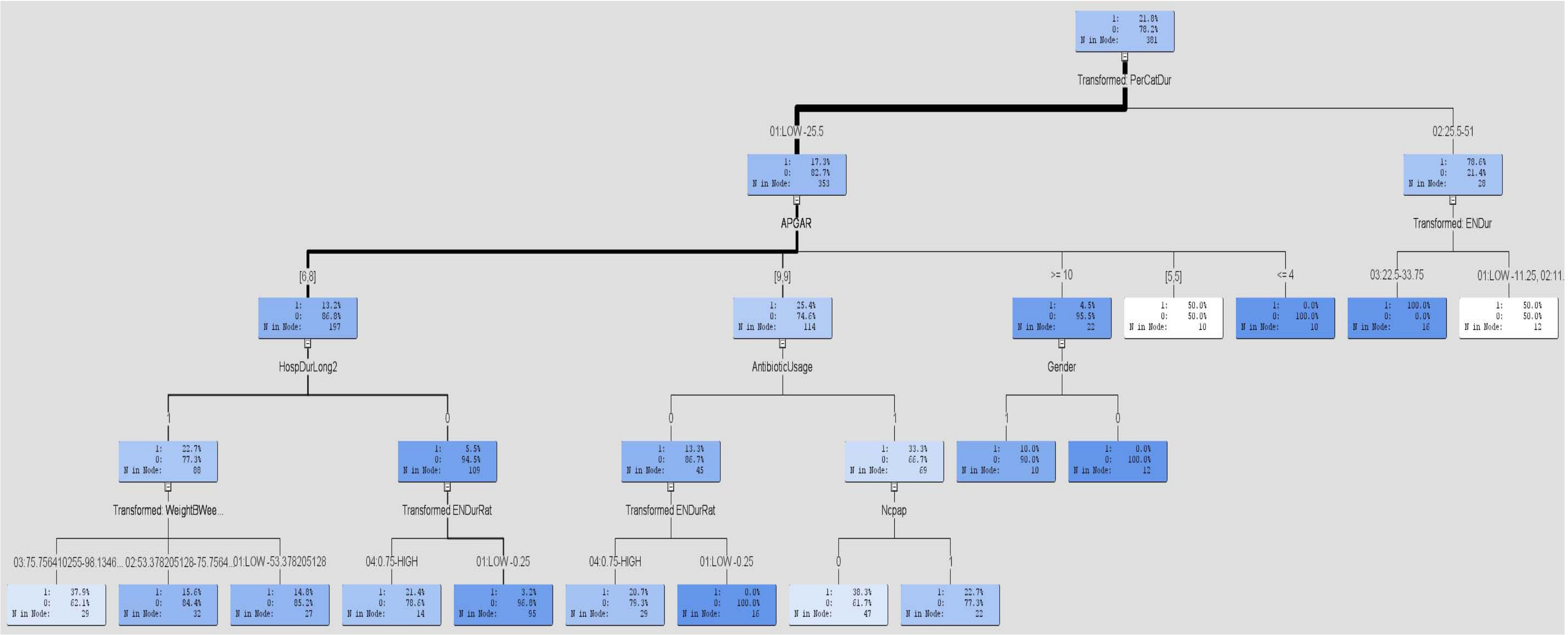

Figure 36. Entropy decision tree with holdout sampling

D. CORRELATION MATRIX

The relationships between independent continuous variables were examined with correlation matrix. For each variable, the first row is the pearson correlation coefficient, the second row is p-value and the third row is the number of observations for which the correlation coefficients were calculated.

Table 41. Correlation Matrix

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations																											
	AgeDay	IntubDurRat	ENDur	Weight	BirthWeek	HospDur	VenDur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPVDur	AspirDur	TPNDur	ICUDur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
AgeDay	1	-0.06614	-0.14531	0.13581	0.14169	-0.07895	-0.06086	-0.07312	0.00583	-0.01353	0.00576	-0.0821	-0.07789	-0.06187	-0.0349	-0.06953	-0.06917	-0.07925	-0.08114	-0.07648	0.13039	-0.07906	-0.08071	0.00682	-0.0225	-0.09764	-0.00474
		0.123	0.0011	0.0015	0.0009	0.0655	0.1559	0.0881	0.892	0.7527	0.8932	0.0554	0.0692	0.1492	0.4161	0.1052	0.1067	0.0645	0.0584	0.0744	0.0023	0.0651	0.06	0.8738	0.6002	0.0282	0.9121
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
IntubDurRat	-0.06614	1	0.14984	-0.2269	-0.30822	0.34003	0.59272	0.37505	-0.00747	-0.01485	0.23471	0.08335	0.64677	0.15774	0.14272	0.50134	0.41354	0.36748	0.43914	0.17814	-0.21164	0.40846	0.63679	0.20502	0.19446	-0.1034	0.37562
	0.123		0.0007	<.0001	<.0001	<.0001	<.0001	<.0001	0.8618	0.7293	<.0001	0.0518	<.0001	0.0002	0.0008	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0201	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	Age Day	IntubDurRat	ENDur	Weight	BirthWeek	HospDur	VenDur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPVDur	AspirDur	TPNDur	ICUDur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
ENDur	-0.14531	0.14984	1	-0.25216	-0.29562	0.51781	0.36022	0.59833	-0.0142	-0.02508	0.14842	0.19468	0.47274	0.41273	-0.06905	0.59699	0.49023	0.36076	0.58622	0.67806	-0.23282	0.22679	0.3974	0.12124	0.30183	0.59202	0.30037
	0.0011	0.0007		<.0001	<.0001	<.0001	<.0001	<.0001	0.7503	0.5739	0.0008	<.0001	<.0001	<.0001	0.1212	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0064	<.0001	<.0001	<.0001
	505	505	505	505	505	505	505	505	505	505	505	505	505	505	505	504	505	505	505	505	505	505	504	505	505	505	505
Weight	0.13581	-0.2269	-0.25216	1	0.81888	-0.46792	-0.21305	-0.45791	-0.08902	-0.06618	0.08082	-0.32698	-0.23927	-0.40462	-0.10079	-0.23395	-0.46003	-0.37104	-0.23451	-0.28038	0.98719	-0.38514	-0.21829	0.09116	-0.1924	0.01569	-0.45335
	0.0015	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	0.0378	0.1228	0.0594	<.0001	<.0001	<.0001	0.0186	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0334	<.0001	0.725	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
BirthWeek	0.14169	-0.30822	-0.29562	0.81888	1	-0.55162	-0.29062	-0.54091	-0.06754	-0.06652	0.05828	-0.42614	-0.33196	-0.47272	-0.19639	-0.31307	-0.56806	-0.4644	-0.3069	-0.3332	0.72717	-0.44857	-0.28189	0.06976	-0.19585	0.00563	-0.51902
	0.0009	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.1153	0.1209	0.1742	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1038	<.0001	0.8996	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
HospDur	-0.07895	0.34003	0.51781	-0.46792	-0.55162	1	0.56624	0.95601	0.21341	0.06407	0.09534	0.43225	0.63626	0.59927	0.21953	0.60571	0.92022	0.73983	0.63505	0.49433	-0.43965	0.49135	0.4468	0.06764	0.19122	-0.05153	0.64872
	0.0655	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.1352	0.026	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1147	<.0001	0.2478	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	Age Day	IntubDurRat	EN Dur	Weight	BirthWeek	HospDur	Ven Dur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPV Dur	AspirDur	TPNDur	ICU Dur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
VenDur	-0.06086	0.59272	0.36022	-0.21305	-0.29062	0.56624	1	0.5918	0.02142	-0.00807	0.16167	0.20157	0.96075	0.14953	0.33063	0.93509	0.64205	0.68138	0.83565	0.12191	-0.20297	0.40454	0.68103	0.13626	0.11925	0.00039	0.3552
	0.1559	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.6177	0.851	0.0002	<.0001	<.0001	0.0005	<.0001	<.0001	<.0001	<.0001	<.0001	0.0044	<.0001	<.0001	<.0001	0.0014	0.0053	0.993	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
PerCatDur	-0.07312	0.37505	0.59833	-0.45791	-0.54091	0.95601	0.5918	1	0.21021	-0.03183	0.10574	0.4463	0.6603	0.58186	0.21151	0.63476	0.91699	0.7513	0.66976	0.50067	-0.43073	0.50215	0.48755	0.0817	0.39137	0.06917	0.64783
	0.0881	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.4583	0.0135	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0566	<.0001	0.1206	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
UmbCatDur	0.00583	-0.00747	-0.0142	-0.08902	-0.06754	0.21341	0.02142	0.21021	1	-0.00936	-0.01645	0.1768	0.01336	0.21859	0.0198	0.02336	0.21681	0.25887	0.02822	0.05995	-0.09596	0.1382	0.01435	-0.02103	0.04961	-0.11101	0.08564
	0.892	0.8618	0.7503	0.0378	0.1153	<.0001	0.6177	<.0001		0.8274	0.7016	<.0001	0.7556	<.0001	0.6447	0.5867	<.0001	<.0001	0.5108	0.1622	0.0251	0.0012	0.7385	0.6242	0.2476	0.0126	0.0457
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
CentCatDur	-0.01353	-0.01485	-0.02508	-0.06618	-0.06652	0.06407	-0.00807	-0.03183	-0.00936	1	-0.00366	-0.02394	-0.00974	0.11057	-0.01159	-0.00744	0.07316	0.08485	-0.00805	-0.00159	-0.06416	0.11135	-0.01215	-0.00484	-0.08098	-0.03577	0.08328
	0.7527	0.7293	0.5739	0.1228	0.1209	0.1352	0.851	0.4583	0.8274		0.9322	0.5771	0.8206	0.0098	0.7872	0.8625	0.0879	0.0477	0.8513	0.9705	0.1347	0.0093	0.7775	0.9103	0.0589	0.4225	0.052
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	Age Day	Intub DurRat	EN Dur	Weight	Birth Week	HospDur	Ven Dur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPV Dur	AspirDur	TPNDur	ICU Dur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
UriCatDur	0.00576	0.23471	0.14842	0.08082	0.05828	0.09534	0.16167	0.10574	-0.01645	-0.00366	1	-0.06025	0.1983	0.00326	-0.00965	0.16529	0.12683	0.10784	0.15563	0.1273	0.07815	0.14206	0.2959	0.9242	0.0533	0.0381	0.16565
	0.8932	<.0001	0.0008	0.0594	0.1742	0.026	0.0002	0.0135	0.7016	0.9322		0.1601	<.0001	0.9395	0.8221	0.0001	0.003	0.0118	0.0003	0.0029	0.0683	0.0009	<.0001	<.0001	0.2141	0.3929	0.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
PhototherDur	-0.0821	0.08335	0.19468	-0.32698	-0.42614	0.43225	0.20157	0.4463	0.1768	-0.02394	-0.06025	1	0.21035	0.35483	0.14007	0.20398	0.43902	0.37173	0.1827	0.26194	-0.28833	0.33487	0.12474	-0.06739	0.1631	-0.01794	0.41082
	0.0554	0.0518	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5771	0.1601		<.0001	<.0001	0.001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0036	0.1161	0.0001	0.6876	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
IntubDur	-0.07789	0.64677	0.47274	-0.23927	-0.33196	0.63626	0.96075	0.6603	0.01336	-0.00974	0.1983	0.21035	1	0.20095	0.3288	0.94968	0.71322	0.68722	0.85681	0.23145	-0.22393	0.42062	0.72237	0.15778	0.13941	0.01622	0.41928
	0.0692	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.7556	0.8206	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0002	0.0011	0.7161	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
NcpapDur	-0.06187	0.15774	0.41273	-0.40462	-0.47272	0.59927	0.14953	0.58186	0.21859	0.11057	0.00326	0.35483	0.20095	1	0.0587	0.19765	0.60459	0.4903	0.21495	0.54474	-0.37141	0.46544	0.23298	0.00372	0.15995	-0.02328	0.57588
	0.1492	0.0002	<.0001	<.0001	<.0001	<.0001	0.0005	<.0001	<.0001	0.0098	0.9395	<.0001	<.0001		0.1712	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.9309	0.0002	0.6017	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	AgeDay	IntubDurRat	ENDur	Weight	BirthWeek	HospDur	VenDur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPVDur	AspirDur	TPNDur	ICUDur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
PPVDur	-0.0349	0.14272	-0.06905	-0.10079	-0.19639	0.21953	0.33063	0.21151	0.0198	-0.01159	-0.00965	0.14007	0.3288	0.0587	1	0.35216	0.2473	0.25259	0.27883	-0.02741	-0.0721	0.11638	0.23382	-0.00528	0.02492	-0.13403	0.14964
	0.4161	0.0008	0.1212	0.0186	<.0001	<.0001	<.0001	<.0001	0.6447	0.7872	0.8221	0.001	<.0001	0.1712		<.0001	<.0001	<.0001	<.0001	0.5231	0.0926	0.0065	<.0001	0.9021	0.5615	0.0025	0.0005
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
AspirDur	-0.06953	0.50134	0.59699	-0.23395	-0.31307	0.60571	0.93509	0.63476	0.02336	-0.00744	0.16529	0.20398	0.94968	0.19765	0.35216	1	0.6914	0.69025	0.89411	0.19334	-0.22065	0.39462	0.77276	0.13895	0.12002	0.11338	0.3637
	0.1052	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5867	0.8625	0.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0012	0.0051	0.0109	<.0001
	544	544	504	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	504	544
TPNDur	-0.06917	0.41354	0.49023	-0.46003	-0.56806	0.92022	0.64205	0.91699	0.21681	0.07316	0.12683	0.43902	0.71322	0.60459	0.2473	0.6914	1	0.81304	0.71421	0.45403	-0.42792	0.56828	0.53994	0.10403	0.20191	-0.04796	0.75969
	0.1067	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0879	0.003	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0151	<.0001	0.2821	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
ICUDur	-0.07925	0.36748	0.36076	-0.37104	-0.4644	0.73983	0.68138	0.7513	0.25887	0.08485	0.10784	0.37173	0.68722	0.4903	0.25259	0.69025	0.81304	1	0.65878	0.30658	-0.34876	0.7438	0.50936	0.07996	0.14204	-0.01781	0.49511
	0.0645	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0477	0.0118	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0621	0.0009	0.6897	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients

Prob > \|r\| under H0: Rho=0

Number of Observations

	AgeDay	IntubDurRat	ENDur	Weight	BirthWeek	HospDur	VenDur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPVDur	AspirDur	TPNDur	ICUDur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
NgUseDur	-0.08114	0.43914	0.58622	-0.23451	-0.3069	0.63505	0.83565	0.66976	0.02822	-0.00805	0.15563	0.1827	0.85681	0.21495	0.27883	0.89411	0.71421	0.65878	1	0.13513	-0.22264	0.36278	0.65427	0.12774	0.13234	0.11017	0.37322
	0.0584	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5108	0.8513	0.0003	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.0016	<.0001	<.0001	<.0001	0.0028	0.002	0.0132	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
OgUseDur	-0.07648	0.17814	0.67806	-0.28038	-0.3332	0.49433	0.12191	0.50067	0.05995	-0.00159	0.1273	0.26194	0.23145	0.54474	-0.02741	0.19334	0.45403	0.30658	0.13513	1	-0.25832	0.27615	0.22733	0.10655	0.15431	0.15739	0.3906
	0.0744	<.0001	<.0001	<.0001	<.0001	<.0001	0.0044	<.0001	0.1622	0.9705	0.0029	<.0001	<.0001	<.0001	0.5231	<.0001	<.0001	<.0001	0.0016		<.0001	<.0001	<.0001	0.0128	0.0003	0.0004	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
WeightBWeekRat	0.13039	-0.21164	-0.23282	0.98719	0.72717	-0.43965	-0.20297	-0.43073	-0.09596	-0.06416	0.07815	-0.28833	-0.22393	-0.37141	-0.0721	-0.22065	-0.42792	-0.34876	-0.22264	-0.25832	1	-0.3566	-0.20375	0.08759	-0.17947	0.02264	-0.41655
	0.0023	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0251	0.1347	0.0683	<.0001	<.0001	<.0001	0.0926	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.0409	<.0001	0.6118	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
ICUDurRat	-0.07906	0.40846	0.22679	-0.38514	-0.44857	0.49135	0.40454	0.50215	0.1382	0.11135	0.14206	0.33487	0.42062	0.46544	0.11638	0.39462	0.56828	0.7438	0.36278	0.27615	-0.3566	1	0.50855	0.10555	0.18095	-0.05995	0.56849
	0.0651	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0012	0.0093	0.0009	<.0001	<.0001	<.0001	0.0065	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.0137	<.0001	0.1786	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	Age Day	Intub DurRat	EN Dur	Weight	Birth Week	HospDur	Ven Dur	PerCatDur	Umb CatDur	Cent CatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPV Dur	AspirDur	TPNDur	ICUDur	NgUseDur	OgUseDur	WeightBWeekRat	ICU DurRat	Aspir DurRat	UriCat DurRat	PerCat DurRat	ENDurRat	TPN DurRat
AspirDurRat	-0.08071	0.63679	0.3974	-0.21829	-0.28189	0.4468	0.68103	0.48755	0.01435	-0.01215	0.2959	0.12474	0.72237	0.23298	0.23382	0.77276	0.53994	0.50936	0.65427	0.22733	-0.20375	0.50855	1	0.27717	0.16983	0.05203	0.43076
	0.06	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.7385	0.7775	<.0001	0.0036	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.2436	<.0001
	544	544	504	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	544	504	544
UriCatDurRat	0.00682	0.20502	0.12124	0.09116	0.06976	0.06764	0.13626	0.0817	-0.02103	-0.00484	0.9242	-0.06739	0.15778	0.00372	-0.00528	0.13895	0.10403	0.07996	0.12774	0.10655	0.08759	0.10555	0.27717	1	0.06421	0.01043	0.16842
	0.8738	<.0001	0.0064	0.0334	0.1038	0.1147	0.0014	0.0566	0.6242	0.9103	<.0001	0.1161	0.0002	0.9309	0.9021	0.0012	0.0151	0.0621	0.0028	0.0128	0.0409	0.0137	<.0001		0.1344	0.8151	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545
PerCatDurRat	-0.0225	0.19446	0.30183	-0.1924	-0.19585	0.19122	0.11925	0.39137	0.04961	-0.08098	0.0533	0.1631	0.13941	0.15995	0.02492	0.12002	0.20191	0.14204	0.13234	0.15431	-0.17947	0.18095	0.16983	0.06421	1	0.36868	0.25905
	0.6002	<.0001	<.0001	<.0001	<.0001	<.0001	0.0053	<.0001	0.2476	0.0589	0.2141	0.0001	0.0011	0.0002	0.5615	0.0051	<.0001	0.0009	0.002	0.0003	<.0001	<.0001	<.0001	0.1344		<.0001	<.0001
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

Table 41. continued.

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	AgeDay	IntubDurRat	ENDur	Weight	BirthWeek	HospDur	VenDur	PerCatDur	UmbCatDur	CentCatDur	UriCatDur	PhototherDur	IntubDur	NcpapDur	PPVDur	AspirDur	TPNDur	ICUDur	NgUseDur	OgUseDur	WeightBWeekRat	ICUDurRat	AspirDurRat	UriCatDurRat	PerCatDurRat	ENDurRat	TPNDurRat
ENDurRat	-0.09764	-0.1034	0.59202	0.01569	0.00563	-0.05153	0.00039	0.06917	-0.11101	-0.03577	0.0381	-0.01794	0.01622	-0.02328	-0.13403	0.11338	-0.04796	-0.01781	0.11017	0.15739	0.02264	-0.05995	0.05203	0.01043	0.36868	1	-0.13136
	0.0282	0.0201	<.0001	0.725	0.8996	0.2478	0.993	0.1206	0.0126	0.4225	0.3929	0.6876	0.7161	0.6017	0.0025	0.0109	0.2821	0.6897	0.0132	0.0004	0.6118	0.1786	0.2436	0.8151	<.0001		0.0031
	505	505	505	505	505	505	505	505	505	505	505	505	505	505	505	504	505	505	505	505	505	505	504	505	505	505	505
TPNDurRat	-0.00474	0.37562	0.30037	-0.45335	-0.51902	0.64872	0.3552	0.64783	0.08564	0.08328	0.16565	0.41082	0.41928	0.57588	0.14964	0.3637	0.75969	0.49511	0.37322	0.3906	-0.41655	0.56849	0.43076	0.16842	0.25905	-0.13136	1
	0.9121	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0457	0.052	0.0001	<.0001	<.0001	<.0001	0.0005	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0031	
	545	545	505	545	545	545	545	545	545	545	545	545	545	545	545	544	545	545	545	545	545	545	544	545	545	505	545

E. PROBABILITY DISTRIBUTION

The distance between cumulative percentage curves indicates the model's classification power. For an accurate model, cumulative percentage of nonevents curve is close to bottom left, cumulative percentage of events curve is close to upper right in the graphic.

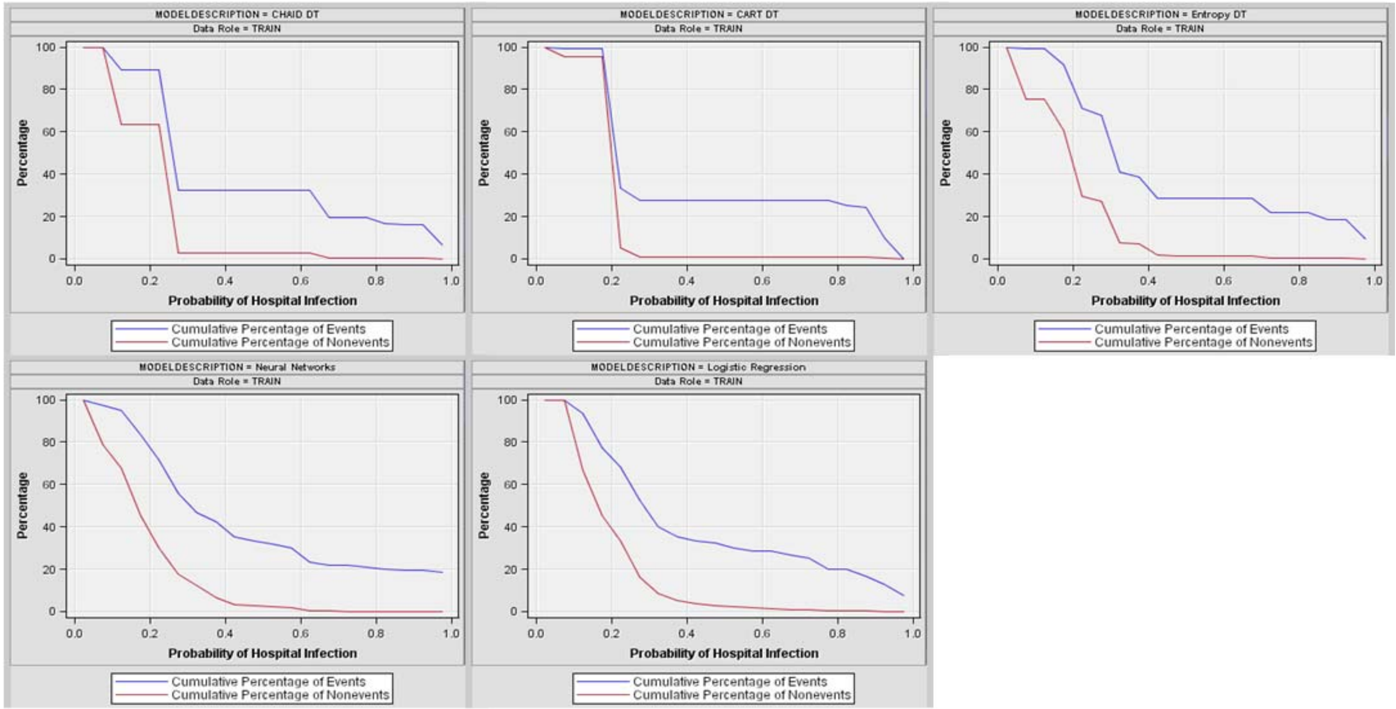

Figure 37. Probability distribution of models

REFERENCES

About CRISP-DM (n.d.) Retrieved June 22, 2008, from: http://www.crisp-dm.org/

Anand, S.S., Buchner, A.G. (1998). *Decision Support Using Data Mining*. London: Financial Times Pitman Publishers

Borghesi, A., Stronati, M. (2008). Strategies for the prevention of hospital-acquired infections in the neonatal intensive care unit. *Journal of Hospital Infection, Volume 68, Issue 4, March 2008 (293-300).*

Box, G.E.P., Cox, D.R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B, 26(2) (211-252).*

Braithwaite, E.A., Dripps, J., Lyon, A.J., Murray, A. (2001) Artificial Neural Networks for Neonatal Intensive Care In Dybowski, R., Gant, V. (Eds.) *Clinical Applications of Artificial Neural Networks* (pp.102-119) Cambridge: Cambridge University Press.

Breaux, D., Baker, J., Wilburn, S., Monteith, D., Umstead, S. (2005). Using automated surveillance to trace evidence-based practices: Reducing infection outcomes when Escherichia coli is your most common uropathogen. *American Journal of Infection Control Volume 33, Issue 5, June 2005.*

Breiman, L. Friedman, J., Olshen, R., Stone, C. (1984) *Classification and Regression Trees.* New York: Chapman & Hall.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1998). *Discovering Data Mining: From Concepts to Implementation.* Saddle River, New Jersey: Prentice Hall.

Chakraborty, B.; Chakraborty, G.; (1999) Rule extraction from structured neural network for pattern classification. *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on* , vol.4, no., pp.869-874 vol.4, 1999

Chen, H., Fuller, S.S., Friedman, C., Hersh, W. (2005). *Medical Informatics Knowledge Management and Data Mining in Biomedicine.* New York: Springer.

Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., Benjamin, Jr. DK. (2004). Nosocomial infection in the NICU: a medical complication or unavoidable problem? *Journal of Perinatology 2004;24: 382-8*. In Borghesi, A., Stronati, M. (2008). Strategies for the prevention of hospital-acquired infections in the neonatal intensive care unit. *Journal of Hospital Infection, Volume 68, Issue 4, March 2008 (293-300).*

CRISP-DM Consortium: Chapman, P. et al. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. Retrieved June 22, 2008, from: http://www.crisp-dm.org/CRISPWP-0800.pdf

Dao, TK. (2006). Soaring to New Heights with Data Mining. *American Journal of Infection Control Volume 34, Issue 5, June 2006,* (Publication Number 21-194).

Dao, T.K., Zabaneh, F., Holmes, J., Disrude, L., Price, M., Gentry, L. (2008) A practical data mining method to link hospital microbiology and an infection control database. *American Journal of Infection Control, Volume 36, Issue 3, Supp.1, April 2008 (18-20).*

Duda, R.O., Hart, P.E., Stork D.G. (2001). *Pattern Classification.* New York: Wiley

Dybowski, R., Roberts, S. (2005) An Anthology of Probabilistic Models for Medical Informatics In Husmeier, D., Dybowski, R., Roberts, S. (Eds.) *Probabilistic Modeling in Bioinformatics and Medical Informatics* (pp.297-349) London: Springer-Verlag.

Erol, S. (2008). Hastane Enfeksiyonları Sürveyansı [Hospital Infections Surveillance]. *Hastane Enfeksiyonları Koruma ve Kontrol Sempozyum Dizisi, 60, (43-51).*

Ertek, M. (2008). Hastane Enfeksiyonları: Türkiye Verileri [Hospital Infections: Statistics of Turkey]. *Hastane Enfeksiyonları Koruma ve Kontrol Sempozyum Dizisi, 60, (9-14).*

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.). (1996). *Advances in Knowledge Discovery and Data Mining.* Cambridge, Massachusetts: The MIT Press

Görenek, L. (2002). *Hastane İnfeksiyonları* [Hospital Infections]. Retrieved June 15, 2008, from http://www.gata.edu.tr/dahilibilimler/infeksiyon/Ders_Notlari.htm

Hair, J.F.Jr., Anderson, R.E., Tatham, R.L., Black, W.C. (1998). *Multivariate Data Analysis, 5/E.* New Jersey: Prentice Hall

Han, J., Kamber, M. (2006). *Data Mining Concepts and Techniques.* San Francisco: Morgan Kaufmann

Husmeier, D., Dybowski, R., Roberts, S. (Eds.). (2005). *Probabilistic Modeling in Bioinformatics and Medical Informatics.* London: Springer-Verlag

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of Applied Statistics, Vol. 29, No. 2, (pp.119-127).*

Kleinbaum, D.G., Klein, M. (2002). *Logistic Regression A Self-Learning Text.* (2nd Ed.) New York: Springer

Kreuze, D. (2001). Debugging Hospitals. *Technology Review; Mar 2001; 104, 2; ABI/INFORM Global pg. 32*

Lamma, E., Manservigi, M., Mello, P., Storari, S., Riguzzi, F. (2000) A System for Monitoring Nosocomial Infections. In Brause, R.W., Hanisch, E. (Eds.), *Medical Data Analysis, ISMDA 2000.* (pp. 282-292) Berlin: Springer

Maimon, O., Rokach, L. (2005). Decision Trees. In O. Maimon, L.Rokach (Eds.), *The Data Mining and Knowledge Discovery Handbook.* (pp. 165-192) New York: Springer Science+Business Media, Inc.

Maimon, O., Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook.* New York: Springer Science+Business Media, Inc.

Matussek, A., Taipalensuu, J., Einemo, I., Tiefenthal, M., Löfgren, S. (2007). Transmission of Staphylococcus aureus from maternity unit staff members to newborns disclosed through spa typing. *American Journal of Infection Control, Volume 35, Issue 2, March 2007 (122-125).*

Meek, J., Tinney, S. (2006). Computerize your infection surveillance for improved patient care-and savings. *Healthcare Financial Management, Volume 60, Issue 12, December 2006 (ABI/INFORM Global pg. 108).*

Perk, Y. (2008). Yenidoğan Yoğun Bakım Enfeksiyonları; Koruma ve Kontrol [Newborn Intensive Care Unit Infections; Protection and Control]. *Hastane Enfeksiyonları Koruma ve Kontrol Sempozyum Dizisi, No:60, (137-141).*

Pittet, D. (2005). Infection control and quality health care in the new millennium. *American Journal of Infection Control, Volume 33, Issue 5, June 2005 (258-67).*

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, Volume 1, Issue1, 1986 (81-106).*

Quinlan, J.R. (1986). *C4.5: Programs for Machine Learning.* San Mateo, California: Morgan Kaufmann Publishers Inc.

Sameshima, H., Ikenoue, T. (2006). Developmental effects on neonatal mortality and subsequent cerebral palsy in infants exposed to intrauterine infection. *Early Human Development, Volume 83, Issue 8, August 2007 (517-9).*

SAS Enterprise Miner, SEMMA. (n.d.). Retrieved June 22, 2008, from: http://www.sas.com/technologies/analytics/datamining/miner/semma.html

SAS Help and Documentation. (n.d.). SAS 9.1 Product Help.

Tan, P.N., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. Boston: Pearson Addison-Wesley

van Rossem, M.C., de Waal, W.J., van Hannen, E.J., Verboon-Maciolek, M.A., van Wieringen, H., Van de Vijver, D.A.M., et al. (2007). Enterobacter colonisation in newborn infants: predictors, follow-up and implications for infection control. *Journal of Hospital Infection, Volume 67, Issue 2, November 2007* (*142-8*).

Usluer, G., Erben, N., Özgüneş, İ., Kartal, E.D., Akşit, F., İhtiyar, E., et al. (2006). A Prospective Study to Evaluate Device-Associated Nosocomial Infection Rates in Intensive Care Units at a University Hospital in Turkey. *Journal of Hospital Infection, Volume 64, Supplement 1, (S98).*

Weiss, S.M., Kulikowski, C.A. (1991). *Computer Systems That Learn.* San Mateo, California: Morgan Kaufmann Publishers Inc.