İstanbul Üniversitesi İşletme Fakültesi Dergisi
Istanbul University Journal of the School of Business Administration
Cilt/Vol:39, Sayı/No:2, 2010, 210-226
ISSN: 1303-1732 – www.ifdergisi.org © 2010

Finding hidden patterns of hospital infections on newborn: A data mining approach

İnci Aksoy[1]
Management Information Systems,
Institute of Social Sciences
Bogazici University, Istanbul, Türkiye

Bertan Badur[2]
Management Information Systems,
School of Applied Disciplines
Bogazici University, Istanbul, Türkiye

Sona Mardikyan[3]
Management Information Systems,
School of Applied Disciplines
Bogazici University, Istanbul, Türkiye

Abstract

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. 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 and neural network classification models are built using accuracy estimation methods; holdout sampling and cross validation. In model comparison, accuracy and sensitivity measures are taken into consideration primarily. The study highlights that antibiotics and urinary catheter usage, peripheral catheter duration, enteral and total parenteral nutrition durations, and birth weight for gestational age are considerable risk factors. Among the models, neural network and CHAID decision tree perform better on hospital infections detection.

Keywords: *Data mining, hospital infections, decision trees, neural networks.*

Hastane enfeksiyonlarının gizli örüntülerinin bulunması: Bir veri madenciliği yaklaşımı

Özet

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. 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. Karar ağaçları ve yapay sinir ağları sınıflandırma modelleri basit ve çapraz doğrulama yöntemleri ile kurulmuştur. Model karşılaştırmada doğruluk ve duyarlılık oranları öncelikli olarak dikkate alınmıştır. Bu çalışmada antibiyotik ve üriner kateter kullanımı, periferik kateter kullanım süresi, enteral ve total parenteral beslenme süreleri ve doğum ağırlığının gestasyonel yaşa oranı önemli risk faktörleri olarak bulunmuştur. Yapay sinir ağları ve CHAID karar ağaçları hastane enfeksiyonlarının tespitinde başarılı olmuştur.

[1] *inciaksoy@yahoo.com (İ. Aksoy)*

[2] *badur@boun.edu.tr (B. Badur)*

[3] *mardikya@boun.edu.tr (S. Mardikyan)*

Anahtar Kelimeler: *Veri madenciliği, hastane enfeksiyonları, karar ağaçları, yapay sinir ağları.*

1. Introduction

Hospital infections or nosocomial infections are infections that originate or occur in a hospital or in a health care service unit [1]. 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.

As reported by Perk [2], 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). 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).

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.

Pittet [3] 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 [4]. The cause of the outbreak was found to be the urinary catheter selection in emergency room. Consequently infection control program was changed and DMSS resulted in an improvement with 3 consecutive months of zero EC urinary isolates.

In order to monitor hospital infections in various areas of the hospital and to identify and report the critical situations for patients, Lamma et al. [5] 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.

The aim of this study is to discover a pattern with prominent reasons of hospital infections on newborn by applying predictive data mining techniques on the data set collected by the Department of Clinical Microbiology and Infectious Diseases in Eskişehir Osmangazi University, Faculty of Medicine. CART (Classification and Regression Trees) and CHAID (Chi-Squared Automatic Interaction Detection) decision trees and neural network data mining techniques are applied. Knowledge Discovery in Databases (KDD)

methodology is followed throughout the study. Holdout sampling and cross validation accuracy estimation methods are applied and models are compared regarding accuracy and sensitivity measures of the test data set. Besides these measures, specificity, area under Receiver Operating Characteristic (ROC) curve, gini coefficient and average squared error measures are taken into consideration.

In the following section, data mining methodology, data mining techniques used in the study and important features of medical data mining are provided. In section 3 data preparation and preprocessing steps are explained. Section 4 includes evaluation of model results with two different accuracy estimation methods. Finally in section 5, the conclusions drawn from the study and possible further research directions are expressed.

2. Methodology

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 methodology (KDD methodology) is used. The process of KDD has been defined by Fayyad et al. [6] as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in 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.

Data mining tasks are generally classified into two broad categories as descriptive and predictive. 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 [7]. Decision trees, neural networks, and logistic regression are examples for classification techniques.

Decision trees are one of the most popular classification and prediction methods used in data mining. Han and Kamber [7] 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 holds a class label. The fundamental principle underlying tree creation is that of simplicity: Decisions that lead to simple, compact tree with few nodes are preferred [8]. The tree complexity is also explicitly controlled by the stopping criteria used and the pruning method employed [9].

In this study, CART and CHAID decision tree algorithms are applied. The CART [10] 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 [11] 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.

Neural 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. Han and Kamber [7] 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.

In this study, multilayer feed-forward network is built using backpropagation algorithm. Feed-forward networks represent non-linear functional mappings between input variables and the output variable [12]. The network consists of one input, one or multiple hidden

layers, and one output layer and nodes are connected only to the nodes in the next layer. Backpropagation algorithm processes each training observation sequentially. 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.

The difference between decision tree and neural network classification techniques can be addressed by their explanation capability. Human can understand the rules induced in by a decision tree and can control the decision making process. On the other hand, a neural network based classifier is more or less a black box.

The predictive models are evaluated with classifier accuracy measures. The accuracy of a classifier on a given test set is the percentage of test set observations that are correctly classified by the classifier [7], whereas the error rate is the percentage of misclassified observations. 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 is used. In medicine, false negatives and false positives are not treated equally, especially in life-threatening illnesses. Therefore, measures like sensitivity and specificity are widely used, which are derived from confusion matrix [13]. Sensitivity (also known as true positive rate) is the proportion of positive observations, whereas specificity (false positive rate) is the proportion of negative observations that are correctly classified. A graphical approach that shows the trade-off between the sensitivity and (1-specificity) for a given model is the ROC curve [7]. ROC curve also allows for comparing the relative performance among several classifiers.

In medical data mining, the desirable features of a general clinical classification model are listed as follows[14]: high accuracy, 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. Besides, sensitivity is of extreme importance as it shows the number of actual occurrences of a condition remain undiagnosed [15].

3. Data Preprocessing

The greatest challenge in handling the data preprocessing 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. All analysis in the study was generated using SAS Enterprise Miner® software[4], Version 5.3 of the SAS System for Unix.

3.1. Data Description

The data set 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. 83 attributes representing those patients were considered in the study. Common characteristics of these attributes are related to patient (gestational age, weight,

[4] Copyright © 2003-2007 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

premature, congenital anomaly etc.), medical problem (meconium aspiration syndrome, acute renal failure, perinatal asphyxia etc.) and treatment (phototherapy, mechanical ventilation, catheter or antibiotic usage etc.). The dependent variable is *HospInfec*. If the newborn is infected, the variable takes the value of "1", otherwise "0".

3.2. Data Cleaning

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, otherwise zero.

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 intensive care unit (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.

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. For 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 1 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 1 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 assigned to the largest branch or used in search for a split where all missing values were treated as having the same unknown non-missing value for continuous variables and as a separate category for categorical variables.

3.3. Data Reduction

Constant variables (variables with a single value) and variables with few distinct values were eliminated from all models. Before applying decision trees, no data reduction technique was used, because data reduction for variable selection is a natural characteristic of decision tree algorithms. 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, a stepwise logistic regression model was built and

the variables selected by this model were used as the input variables in neural networks model.

3.4. Data Transformation

Variables were transformed with different methods before applying different classification techniques. Before CHAID decision tree 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.

Before neural networks, "equalize spread by target variable" method, which is a subset of Box-Cox transformations [16], 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.

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

Thus the variance in the interval variables between different levels of target is stabilized. Table 2 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 2 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 (*ENDur*) and peripheral catheter duration (*PerCatDur*) 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.

4. Evaluation of Model Results

In this study, CHAID and CART decision tree and neural network models are built. The same accuracy estimation methods were used for all models. First accuracy estimation method was selected to be holdout stratified sampling with typical ratio of 70% training and 30% test samples [17]. Second accuracy estimation method was 10-fold cross validation and the parameters of classification algorithms decided with holdout method were adopted. This approach facilitated parameter selection and detailed analyses of model results. In 10-fold cross validation method, 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 performed according to the test sample results of this method.

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.

4.1. CHAID Decision Tree Model

CHAID decision tree model was built with the parameters illustrated in Table 3. 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 3 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

4.1.1. CHAID Decision Tree Model with Holdout Sampling

CHAID decision tree model selected *AntibioticUsage* (Antibiotics Usage)*, ENDur* (Enteral Nutrition Duration)*, PerCatDur* (Peripheral Catheter Usage Duration)*,* and *UrinCatheter* (Urinary Catheter Usage) variables with holdout sampling. All variables of the model are related to medical treatments. Analyzing the effects of the variables, it is seen that as enteral nutrition and peripheral catheter durations 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. The goodness of fit statistics and ROC curves for training and test samples are given in Table 4 and Figure 1 respectively.

Table 4 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.

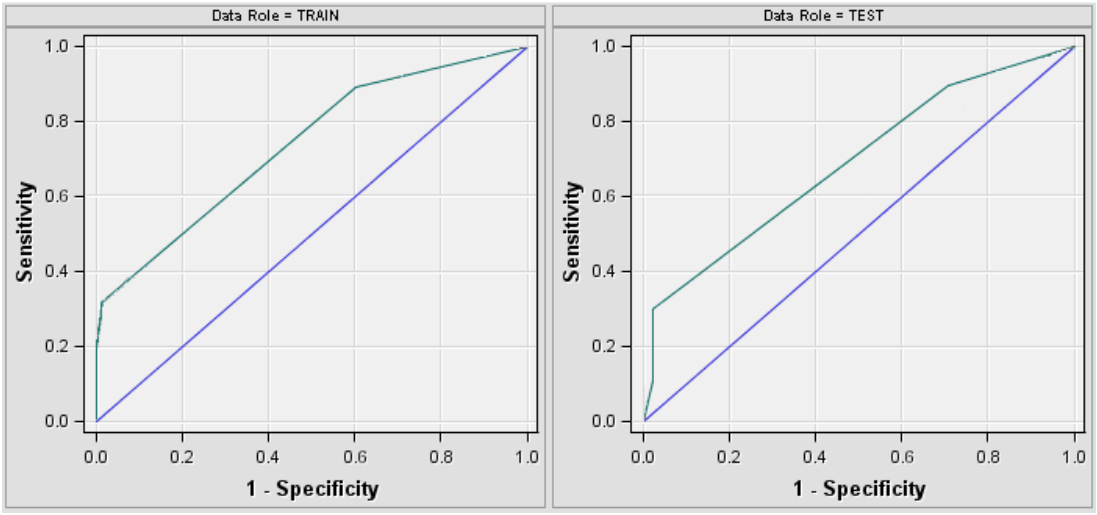

Figure 1 CHAID Holdout Model ROC Curves

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

4.1.2. CHAID Decision Tree Model with Cross Validation

The CHAID decision tree was trained with 10-fold cross validation using the predefined parameters in holdout method. Variables *AntibioticUsage, ENDurRat* (Enteral Nutrition Duration Ratio)*, EnteralNut* (Enteral Nutrition)*, HospDurLong2* (Long Hospital Duration)*, LowAPGAR* (Low APGAR Score)*, PerCatDur* (Peripheral Catheter Duration)*, TPNDur* (Total Parenteral Nutrition Duration)*,* and *UrinCatheter* were selected by CHAID models. It is seen that beside 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 5 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 it is close to the error rate of holdout sampling.

Table 5 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 2 illustrates the ROC curve provided by the training folds in cross validation.

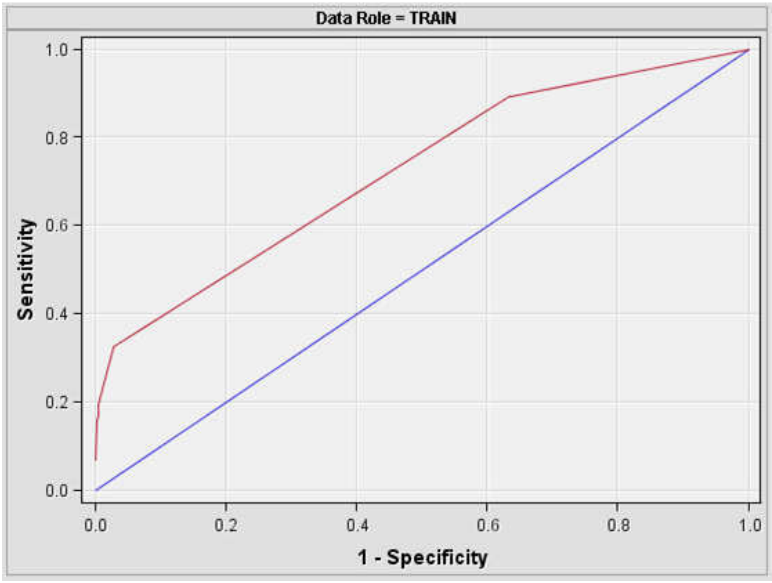

Figure 2 CHAID Cross Validation Model ROC Curve

4.2. 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.

CART algorithm performs cost-complexity pruning [10] by comparing the average squared error between training and validation samples. In addition to pruning, the tree growth was restricted with the stopping rules shown in Table 6.

Table 6 CART Decision Tree Parameters

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

4.2.1. CART Decision Tree Model with Holdout Sampling

Variables *TPNDur* and *PerCatDur* were selected by CART decision tree model with holdout sampling. According to the classification rules represented by model, the probability of hospital infections increases, as total parenteral nutrition duration and peripheral catheter duration 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 7 and Figure 3 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 7 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

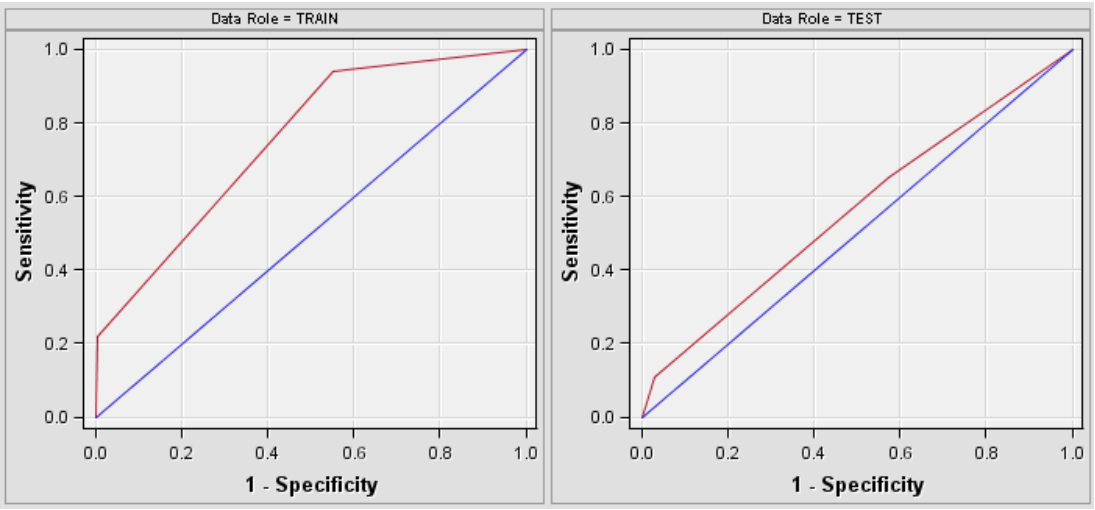

Figure 3 CART Holdout Model ROC Curves

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

4.2.2. CART Decision Tree Model with Cross Validation

Seven variables were selected by the ten CART decision tree models built with cross validation method. These are *AspirDurRat* (Aspiration Duration Ratio)*, ENDurRat, PerCatDur, PerCatDurRat, TPNDur, UriCatDur,* and *Weight* (Birth Weight of the Newborn).

Goodness of fit statistics provided by averaging the training and test folds are represented in Table 8. In comparison to holdout model, cross validation results in training and test samples are more consistent.

Table 8 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 4. The figure shows an improvement in sensitivity measure at lower values of "1-Specificity" in comparison to holdout model.

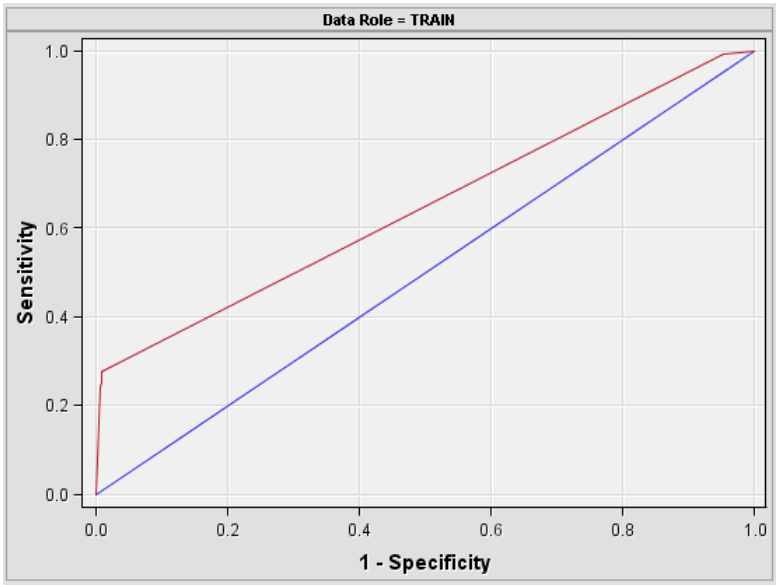

Figure 4 CART Cross Validation Model ROC Curves

4.3. Neural Network Model

In this study, a multilayer feed-forward network 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 and logistic activation function were used in hidden and output layers respectively. 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 Table 9.

Table 9 Neural Networks Parameters

Parameter	Value
Number of Hidden Units	4
Maximum Iterations	100

As explained in section 3, variables *ENDur*, *PerCatDur*, *UriCatDurRat* (Urinary Catheter Duration Ratio), and *UrinCatheter* which 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. Continuous input variables *ENDur*, *PerCatDur* and *UriCatDurRat* were transformed and categorical variable *UrinCatheter* was recoded from {0,1} to {-1,1} in order to converge a good local optimum where the error rate is lower.

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.

4.3.1. Neural Networks Model with Holdout Sampling

Using predefined parameters and determined input variables, neural network model with holdout sampling was built. The goodness of fit statistics represented in Table 10 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 10 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

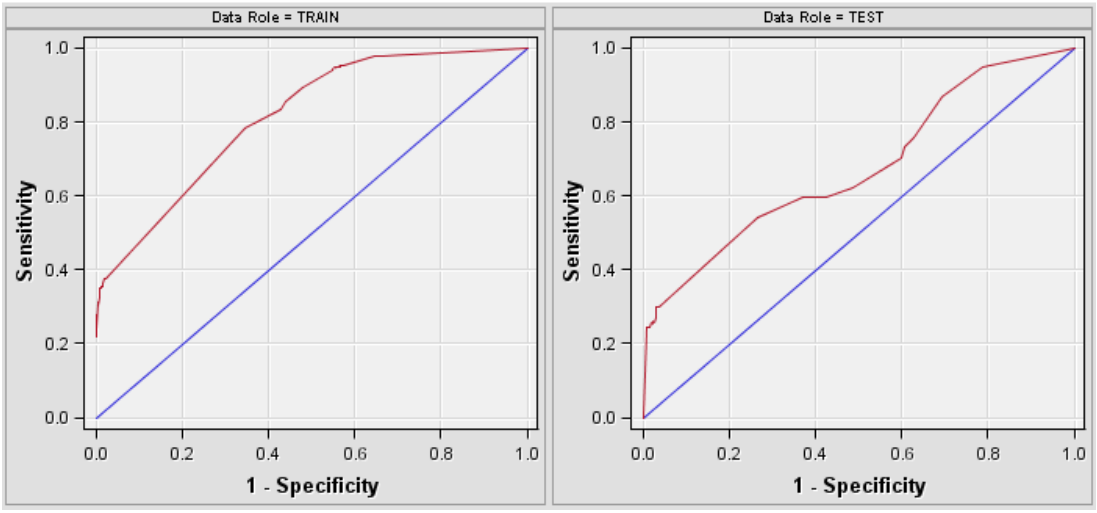

Figure 5 Neural Networks Holdout Model ROC Curve

The decrease in sensitivity and specificity measures of test sample drew down the ROC curve as illustrated in Figure 5.

4.3.2. 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 11 and Figure 6 respectively.

Table 11 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

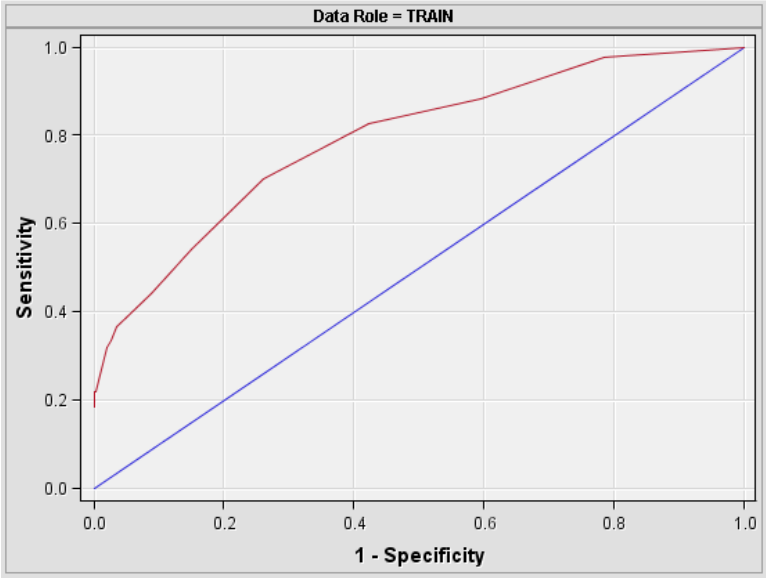

Figure 6 Neural Networks Cross Validation Model ROC Curve

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.

4.4. 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 this study.

There are several criteria that can be used to compare the performance of classification models. While assessing the performance of medical data mining models, the error rate, accuracy and sensitivity measures have higher importance compared to other measures such as root mean squared error. In this 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 results, the model with lowest error rate, highest accuracy and sensitivity is CHAID decision tree. The second model is neural network. The training sample goodness of fit statistics and the ROC curve provided by training folds are given in Table 12 and Figure 7 respectively.

Table 12 Training Sample Goodness of Fit Statistics by Model

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

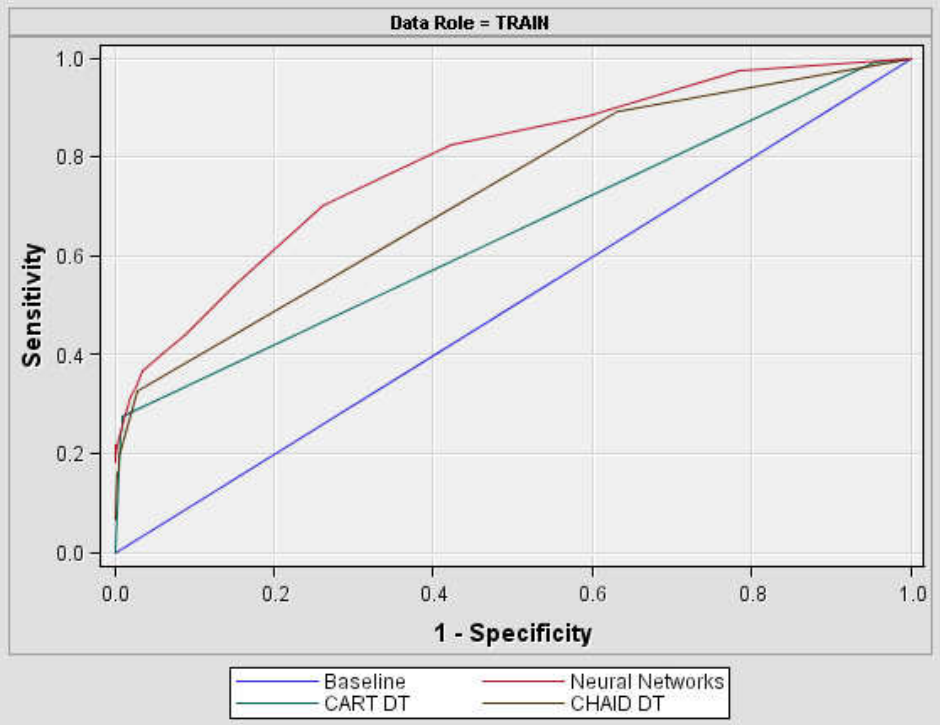

Figure 7 Model Comparison Cross Validation ROC Curves

According to test sample results, there are two successful models in terms of selected 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 Table 13.

Table 13 Test Sample Goodness of Fit Statistics by Model

Statistics	CHAID	CART	Neural Network
Average Squared Error	0.14	0.15	0.14
Area under ROC	0.68	0.65	0.74
Gini Coefficient	0.36	0.31	0.48
Error Rate	0.17	0.19	0.17
Accuracy	0.83	0.81	0.83
Sensitivity	0.30	0.26	0.30
Specificity	0.98	0.97	0.97

According to test ROC and Gini coefficient measures, neural network model is found to be the most accurate model. 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.

5. Conclusion

The surveillance of hospital infections in newborn ICUs has an extreme importance to prevent the outbreaks. 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 this study, in order to discover a pattern with prominent reasons of hospital infections on newborn KDD methodology is used. First, data cleaning and preprocessing steps are applied. CHAID and CART decision trees and neural network models are built in modeling step. Because of the small sample size, in addition to holdout sampling, models are built and compared with cross validation.

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

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 further studies, more observations regarding hospital infections in newborn ICUs should be analyzed with classification algorithms. In addition, the sample should be collected from several hospitals to be able to eliminate the local effects.

References

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

[2] Y. Perk, Yenidoğan Yoğun Bakım Enfeksiyonları; Koruma ve Kontrol. *Hastane Enfeksiyonları Koruma ve Kontrol Sempozyum Dizisi*. 60, 137-141 (2008).

[3] D. Pittet, Infection Control and Quality Health Care in the New Millennium. *American Journal of Infection Control*. 33, 5, June, 258-267 (2005).

[4] D. Breaux, et al., Using Automated Surveillance to Trace Evidence-Based Practices: Reducing Infection Outcomes when Escherichia Coli is Your Most Common Uropathogen. *American Journal of Infection Control*. 33, 5, June, (2005).

[5] E. Lamma, et al., A System for Monitoring Nosocomial Infections, in *Medical Data Analysis ISMDA 2000* (Brause, R.W., Hanisch, E. Eds.). Springer, Berlin, 2000.

[6] U.M. Fayyad, et al. (Eds.), *Advances in Knowledge Discovery and Data Mining.* The MIT Press, Cambridge, Massachusetts, 1996, p.6.

[7] J. Han, Kamber, M., *Data Mining Concepts and Techniques.* Morgan Kaufmann, San Francisco, 2006, p.24, 291, 327, 360, 372.

[8] R.O. Duda, et al., *Pattern Classification.* Wiley, New York, 2001, p.398.

[9] O. Maimon, L. Rokach, Decision trees, in *The Data Mining and Knowledge Discovery Handbook* (O. Maimon, L.Rokach Eds.), Springer Science+Business Media Inc., New York, 2005.

[10] L. Breiman, et al., *Classification and Regression Trees*. Chapman & Hall, New York, 1984, p.30, 66.

[11] G.V. Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of Applied Statistics.* 29, 2, 119-127 (1980).

[12] C.M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995, p.117, 140.

[13] P.N. Tan, et al., *Introduction to Data Mining*. Pearson Addison-Wesley, Boston, 2006, p.296.

[14] R. Dybowski, S. Roberts, An anthology of probabilistic models for medical informatics, in probabilistic modelin*g in Bioinformatics and Medical Informatics.* (Husmeier, D., Dybowski, R., Roberts, S. Eds.), Springer-Verlag, London, 2005.

[15] E.A. Braithwaite, et al., Artificial Neural Networks for Neonatal Intensive Care, in *Clinical Applications of Artificial Neural Networks* (Dybowski, R., Gant, V. Eds.), Cambridge University Press, Cambridge, 2001.

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

[17] S.M. Weiss, C.A. Kulikowski, *Computer Systems That Learn*. Morgan Kaufmann Publishers Inc., San Mateo, California, 1991, p.30.