Applying Data Mining Technique for Predicting Incubator Length of Stay in Egypt and USA

Hagar Fady S.
Dept. Of Computer Science & Eng.,
Faculty of Electronic Engineering
Menoufiya University, Egypt

Taha EL-Sayed T.
Dept. Of Electronics& Electrical communications,
Faculty of Electronic Engineering
Menoufiya University, Egypt

Mervat Mahmoud M.
Dept. Of Computer Science & Eng.,
Faculty of Electronic Engineering
Menoufiya University, Egypt

ABSTRACT

This research aims to provide intelligent tool to predict incubator Length of Stay (LOS) of infants in Egypt and USA, in addition to define factors affecting Length of Stay (LOS) of preterm infants and their severity of impact.

This study has relied on data collected from both Egypt and US, to compare LOS's driving factors and the accuracy of LOS prediction in both countries.

The obtained results indicated that accurate and early diagnosis may speed up treatment and thus reduce length of stay.

General Terms

Data mining, Infant Incubator, Length of Stay

Keywords:Length of Stay, Data Mining, Regression, Incubator, Premature.

1. INTRODUCTION

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst[1]. In healthcare, data mining is becoming increasingly popular, if not increasingly essential [2].

Predictive models in clinical medicine are 'tools for helping decision making that combine two or more items of patient data to predict clinical outcomes'. Such models may be used in several clinical contexts by clinicians and may allow a prompt reaction to unfavorable situations[1].

According to World Health Organization (WHO), newborn deaths, that is deaths in the first four weeks of life (neonatal period), today account for 41% of all child deaths before the age of five. The first week of life is the riskiest week for newborns, and yet many countries are only just beginning postnatal care programsto reach mothers and babies at this critical time.Almost 99% of newborn deaths occur in the developing world. With a reduction of 1% per year, Africa has seen the slowest progress of any region in the world.Existing interventions can prevent two-thirds or more of these deaths if they reach those in need [3].

A neonatal intensive care unit, usually shortened NICU (pronounced "Nickyoo"), is a unit of a hospital specializing in the care of ill or premature newborn infants. Infants are cared for in incubators or "open warmers" [4].Critical care providers are faced with resource shortages including beds to holdadmitted patients. This resource constraint is particularly important in specialized areas of the hospital, suchas intensive care units (ICU) or step down units.An early and accurate prognosis of LOS may have organizational, economic, and medical implications. At times of reduced health care budgets, optimal resource planning, e.g. staff scheduling and early discharge policy, is vital[5]. Evaluating LOS informationis a challenging task , but is essential forthe operational success of a hospital. Intensive careresources in particular are often limited and posescheduling problems for hospital staff and administrators. Predicting LOS is difficult andoften only done retrospectively [2].

Major contributions of this research are:

1. Studying LOS through deploying data mining technique with different algorithms.
2. Comparative study, based on real data collected, between Egypt (African developed country) and USA.This Comparative study casted light on differences in factors and patterns affecting LOS.

This paper is organized as follows. Section 2, reviews related work in LOS prediction. Section 3demonstrates process, algorithms and structure of module used in LOS prediction. Section 4, describes the research results. Section 5, conclusion and key findings of the research

2. RELATED WORK

Few literatures areaddressing LOS prediction in the high-risk patient population of extremely preterminfants. These studies have focused on the effects of specific morbidities on LOS or explored variables that were associated with a pre-specified LOS[6]. Ref.[5], predictedLOS for preterm neonates using multiple linear regression model (MR) and an artificial neural network (ANN) based on few prenatal, perinatal and neonatal factors.Ref. [7],used data mining to predict length of stay in a geriatric hospital department. They applied one of the two classifiers: decision tree C4.5 and its successor R-C4.5s, Naïve Bayesian classifier (NBC) and itssuccessor NBCs. Besides, Naive Bayesian imputation (NBI) model is used for handling missing data.In 2009, linear and logistic regression models with time dependent covariate inclusion were developed by Hintz et al [6]. These models were designed to predict LOS as continuous and categorical variable for infants <27 weeks estimated gestational age.

3. RESEARCH METHODOLOGY

This research consists of three main stages as shown in Fig 3 below:

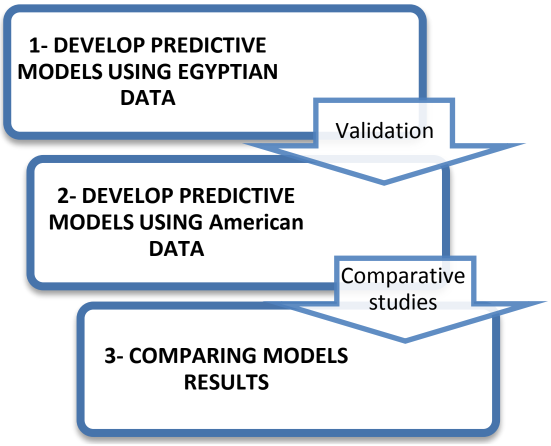

1- DEVELOP PREDICTIVE MODELS USING EGYPTIAN DATA
Validation
2- DEVELOP PREDICTIVE MODELS USING American DATA
Comparative studies
3- COMPARING MODELS RESULTS

Fig. 1 Research methodology

CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology is used as the process to develop data mining modules. This methodology consists of six phases (Business understanding, Data understanding, Data preparation, Modeling, Evaluation, Deployment) intended as a cyclical process (see Fig. 2.) [8].

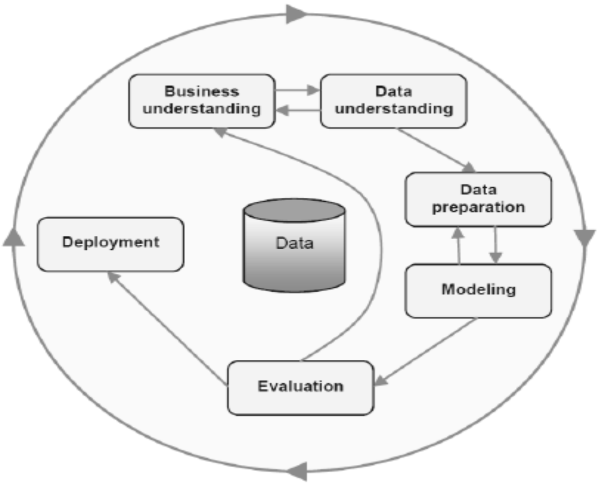


Fig. 2 CRISP-DMprocess

3.1Business and Data Understanding

3.1.1 Procedures and Forms of EGNN,NHDS:

EGNNis a not-for-profit organization whose mission is to improve the efficacy and efficiency of neonatal care in Egypt through a co-ordinate program of research, education, and quality improvement[9]. It has published “28 Day/Discharge Form” and “EGNN_Dataset_Definition_Manual_2010” which is being used in Egyptian neonatal units, in order to standardize data recorded for preterm infants.

NHDS, The National Hospital Discharge Survey, covers discharges from non-institutional hospitals, exclusive of Federal, military, and Veterans Administration hospitals, located in the 50 States and the District of Columbia [10]. Medical information is coded using the “International Classification of Diseases, 9th revision, Clinical modification (ICD-9-CM)” [11].

3.1.2 Previous research

It was found that majority of previous researches developed LOS prediction models based on few risk factors. Those studies used statistical techniques such as multiple linear regression, logistic regression and neural network in LOS prediction model. It also focused on studying LOS in one country\region, rather than analyzing how LOS pattern changes from a region to the other.

3.2Data Preparation

Data preparation includes 3 preprocessing steps(data selection, data cleaning and data transformation) applied to the data to help improve the accuracy,efficiency, and scalability of the classification or prediction process.

3.2.1 Data Selection

The study is carried on:

- Infants admitted to neonatal care unit .Tanta University Hospital from December 2010 and December 2011. Number of records collected 302. Cases were organized as per EGNN form and guidelines [12][13].
- Data available by NHDS “National Hospital Discharge Survey, 2002”[10]. Number of casesstudied 5000. Cases were organized as per NHD description, while diseases (diagnoses) were coded as per “International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)”[11] issued by Center of disease Control and Prevention (CDC).

Dead infants and Cases transferred to other hospitals before being discharged home were excluded. Data were organized and stored in electronic format, guided by the guidelines of EGNN, NHD and advice of preterm expert.

Table1.Sample offactors studied in this research

#	Factor name	Attribute Name
1	Admission number	Admission_no
2	Gestational age	M_Age
3	Birth Weight	Birth Weight
4	Prenatal Care	Prenatal_care
5	Multiple Birth	Multiple_Birth
6	Apgar 1Min __ __ Apgar 5 Min	- Start_Apgar -End_Apgar
7	Respiratory Support After Leaving Delivery Room(a,b,c,d,e,f)	- respiratory_supp_ a/b/c/e
8	Steroids For CLD	Steroids
9	Indomethacin/Ibuprofen For PDA (Prophylactic)	Indomethacin1
10	Addicted mother	Addicted_mother
11	Surgery	Surgery
12	(RDS) Respiratory Distress Syndrome	RDS
13	focal gestational perforation	FGP
14	Cystic Periventricular Leukomalacia	Leukomalacia
15	Hypoxic Ischemic Encephalopathy (HIE)	HIE
16	Patent Ductus Arteriosus (PDA)	PDA
17	Jaundice	Jaundice
18	Maxim.t.bilirubin	Maxim
19	Procedures	Procedures
20	Length Of Stay	LOS/ Dayscare
21	Blood disorder	Blood_disorder
22	Length of stay as category variable	LOS_CAT

3.2.2 DataCleaning

In order to perform LOS prediction using Oracle Data Miner,a column addressing the source of the data was added to the master table.

It is found that in some cases collected from Egypt, data wasincomplete, noisy and inconsistent. In this step,filling on missing values, smoothing out noise while identifying outliers and correcting inconsistencies in the data were performed. Historical Data collected.

3. 2.3Data Transformation

The study sample was randomly splitas follow:

Table2. Data Split

Source of Data	Development	Test	Total
Egypt	260	92	302
USA	3500	1500	5000

Data were discretized (that is, binned); numerical data binned into ranges of values (Quantile binning strategy), and categorical data divided into one bin for each of the values with highest distribution (TopN strategy) and the rest recoded into a bin named “Other”.

3.3Module Development(modeling)

3.3.1Tool Selection

The following tools are used in this research:

I – Oracle Data Miner version11.1.0.4, for the mining activity, that acts as a client and 11g database release 11.1.0.6.0 as a server.

II – Oracle SQL developer.

III – MS Excel.

3. 3.2 Defining Strategies and Algorithms

The following flow chart (Fig. 3) shows the work strategies and algorithms:

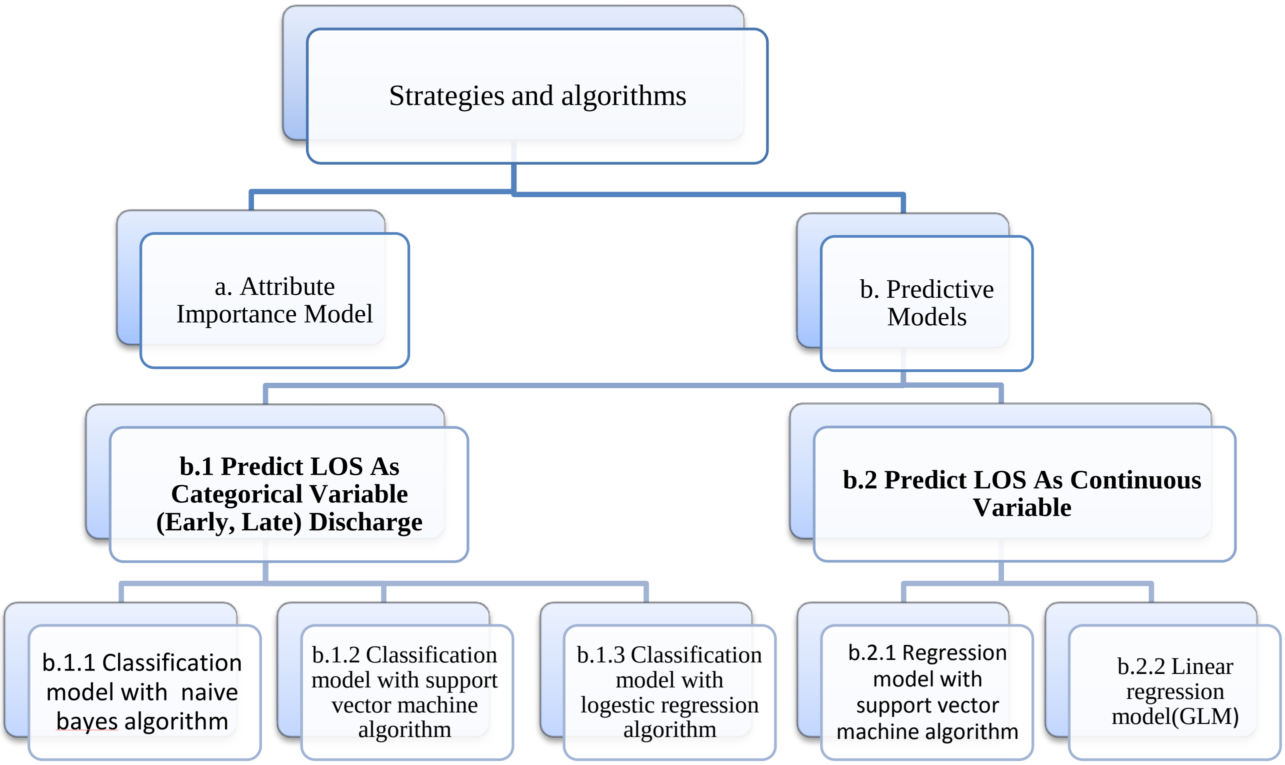

Strategies and algorithms
a. Attribute Importance Model
b. Predictive Models
b.1 Predict LOS As Categorical Variable (Early, Late) Discharge
b.2 Predict LOS As Continuous Variable
b.1.1 Classification model with naive bayes algorithm
b.1.2 Classification model with support vector machine algorithm
b.1.3 Classification model with logestic regression algorithm
b.2.1 Regression model with support vector machine algorithm
b.2.2 Linear regression model(GLM)

Fig. 3Working strategies and algorithms

a. AttributeImportance Model

Oracle Data Mining provides a feature called Attribute Importance (AI) that uses the algorithm Minimum Description Length (MDL) to rank the attributes by significance in determining the target value. Attribute Importance can be used to reduce the size of a classification problem, giving the user the knowledge needed to eliminate some attributes, thus increasing speed and accuracy[14].

b. Predictive Model

b.1 Predict LOS as Categorical Variable (early, late)Discharge

Three predictive models were constructed using classification method with Naive Bayes,SVM and logistic regression algorithms.

b.2 Predict LOS as Continuous Variable(the Number of Days Spent at Incubator)

2 predictive models were constructed using regressionmethod with support vector machine and linear regression algorithm to predict the target variable (LOS).

In module deployment phase, Simple user interfaces were designed to enable medical team to operate the module in an efficient way without perquisite of IT\programming background.

4. RESEARCH RESULTS

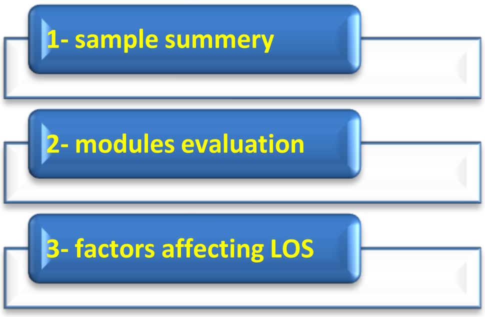


Fig.4 Research results

4.1Sample Summary

4.1.1 Sample summary for EGNN models

1) 78.81% of infants were categorized as early discharge and 21.19% were late.

2) Besides prematurity and neonatal jaundice, common diseases were included in the sample i.e. sepsis, Infant respiratory distress syndrome (RDS) due to immaturity of the lungs.

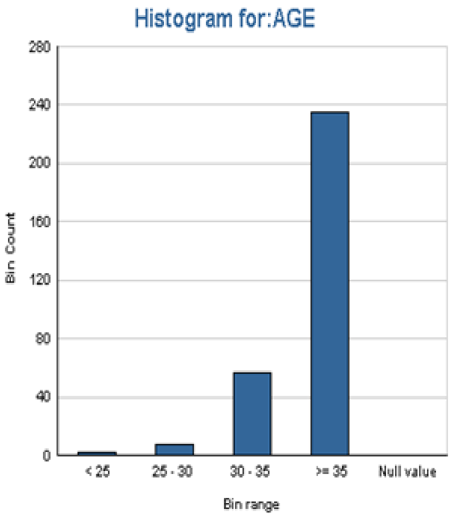


Fig. 4Sample distribution – Age

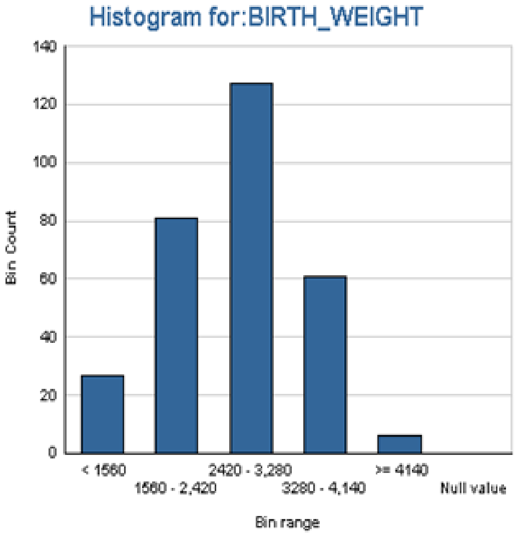


Fig. 5 Sample Distribution – Birth Weight

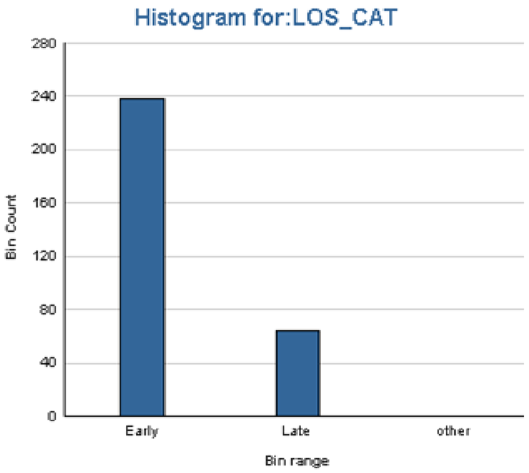


Fig. 6 Sample distribution – LOS – Category

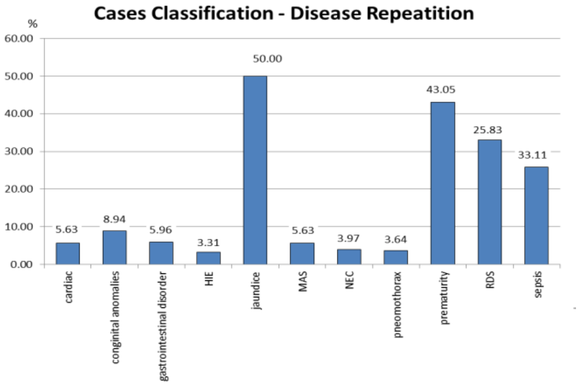
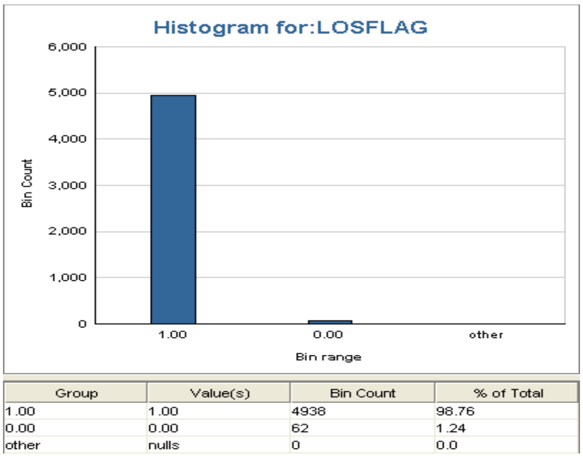


Fig. 7Diseases frequency in EGNN models

4.1.2 Sample summary for NHDS models

1) 54.9% of infants were male and 45.1% were female.
2) 1.24% of infants were dischargedafter 1 day (losflag =0) (early) and 98.76% were late(losflag = 1).
3) Necrotizing Enterocolitis (NEC), prematurity, sepsis and neonatal jaundice takes a highest percentage in diseases frequency.



Group	Value(s)	Bin Count	% of Total
1.00	1.00	4938	98.76
0.00	0.00	62	1.24
other	nulls	0	0.0

Fig. 8Sample Distribution – NHDS

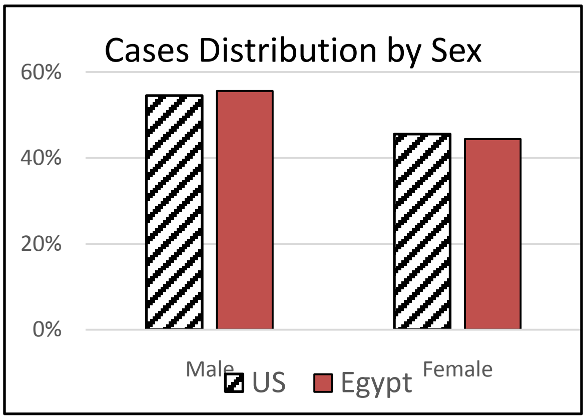


Fig.9Cases distribution by Sex in USA, Egypt

Table3. Diseases Distribution in USA,Egypt

Factor	US	Egypt
Addicted Mother	10%	0%
Blood Disorder	3%	0%
Cardiac	7%	5.63%
Congenital anomalies	17%	9%
gastrointestinal disorder	4%	5.96%
IUGR	22%	2%
Jaundice	41%	58%
NEC	43%	3.97%
prematurity	21%	13%
RDS	16%	40%
sepsis	20%	26%

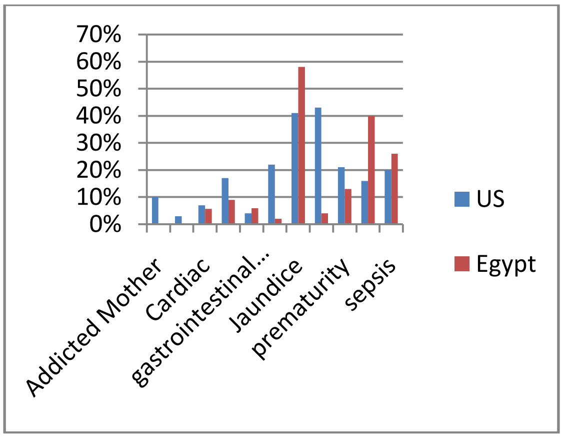


Fig. 10 Diseases frequency

4.2 MODULES EVALUATION

Module validation and testing was carried independentlyfor EGNN and NHDS data. Results are compared in tables 6 and 7. Predictive confidence is a visualindication of the effectiveness of the model compared to a guess based on the distribution of target values in the build dataset [14]. If the model has a predictive confidence of 65.64% that means it is 65.64% better than naïve model.

The modules performance was also compared for EGNN and NHDS. (See table 4, 5)

The comparative tables (4, 5) of predictive capabilities in case of algorithms used in our study and previous research's algorithm presented by Hintz et al [6] For EGNN and NHDS shows that:

1- LOS prediction models as categorical variable

For EGNN and NHDS models, the support vector machine algorithm outperforms others as it presented high predictive confidence, average accuracy and Area Under Curve (AUC) followed by naïve bayes algorithm. On other hand, the logistic regression algorithm presented poor capabilities compared with our algorithms.

2- LOS prediction models as continuous variable

In EGNN and NHDS models, the SVM regression algorithm presented higher capability than linear regression algorithm by means of lower error and higher predictive confidence. SVM are becoming increasingly popular in medicine [15]. It can emulate some traditional methods, such as linear regression and neural nets, but goes far beyond those methods in flexibility, scalability, and speed [14].

By comparing EGNN and NHDS modules performance (see Fig.11, Fig.12), it was observed that all algorithms have close predictive confidence in both EGNN and NHDS data sets, excepts for logistic regression which showed poor predictive confidence with NHDS data set.

LR (Logistic Regression) is not widely used for data mining because of an assumption that LR is unsuitably slow for high –dimensional problems[16].

Overall, the predictive validity of the research models (EGNN AND NHDS) was very good to excellent, with point estimates for the AUC of the Receiver Operating Characteristic(ROC) curves ranging between 0.91 and 0.89.

4.2.1 Evaluation of classification models

Table4. Performance indicator comparison for categorical variable

Performance indicator		support vector machine	naïve bayes	logistic regression *
Predictive confidence	Egypt	0.71	0.53	0.41
	US	0.74	0.57	0.11
Average accuracy	Egypt	0.85	0.76	0.70
	US	0.87	0.78	0.55
Overall accuracy	Egypt	0.89	0.82	0.83
	US	0.79	0.82	0.98
Area under ROC curve	Egypt	0.91	0.87	0.84
	US	0.90	0.86	0.87

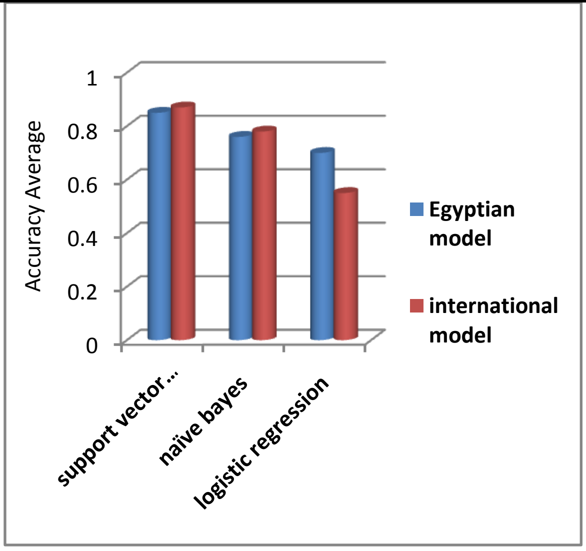


Fig. 11comparing models accuracy in Egypt and USA

4.2.2 Evaluation of Regression models

Table5.Performance indicator comparison for continuous variable

Performance Indicator		support vector machine	Linear regression*
Predictive confidence	Egypt	38.52	22.02
	US	24.54	15.99
Mean absolute error	Egypt	5.65	10.44
	US	2.49	3.56
Root Mean Square Error	Egypt	11.76	14.9
	US	7.38	8.21

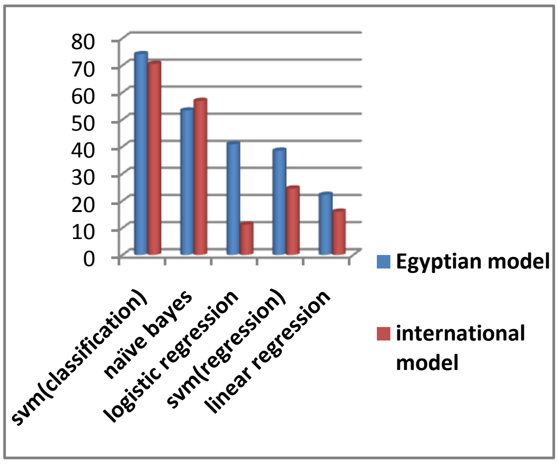


Fig. 12Comparing models predictive confidence in Egypt and USA

4.3 factors affecting LOS

Table 6, 7 demonstrates risk factors highly influencing LOS in Egyptian and American environment. Steroids,sepsis,FGP (Focal GastrointestinalPerforation) and Down syndrome were highly influencing LOS prediction process in EGNN forms. On other hand, leukomalacia, meningitis, ROP (Retinopathy of Prematurity), blood disorder and RDS were effective in NHDS.

Table6.Factorshighly affecting LOS in EGNN model

factor	Attribute importance	Naïve bayes	Logistic regression	Linear regression
BPD		*		*
Down syndrome		*	*	*
FGP	*	*		
IDM			*	*
Respiratory support	*			*
Sepsis	*		*	
Steroids	*	*	*	*
Surgery	*			*

Table7. Factors highly affecting LOS in US model

Factor	Attribute importance	Naïve bayes	Logistic regression	Linear regression
Blood disorder			*	*
Down syndrome		*	*	
Endocrinal		*	*	*
HIE		*	*	
leukomalacia		*	*	*
MAS		*	*	
meningitis		*	*	*
Metabolic acidosis			*	*
Neurological		*	*	
Prematurity	*			*
RDS	*			*
ROP		*	*	*

Table8. Factors affecting LOS in US and EGNN model

Factor	Found in NHDS	Found in EGNN	Effective in NHDS	Effective in EGNN
Steroids		*		*
Sepsis	*	*	*	*
MAS	*	*	*	
Blood disorder	*		*	
RDS	*	*	*	
Prematurity	*	*	*	*

The incidence of the listed risk factors was associated with increasing length of stay. Some factors were found in EGNN, NHDS data sets but highly impacting LOS in one environment. (See table8). Respiratory Distress Syndrome (RDS) is one of them, it was not an effective factor in LOS prediction in EGNN while it highly affect in NHDS. This is due to inaccurately defining it as it was defined as any degree of respiratory impairment ranging from very mild conditions to severely affected cases.

NHDS were more meticulous as they include only sever forms of respiratory disorder.

MAS (Meconium Aspiration Syndrome) is available in both EGNN, NHDS models but was more effective in LOS in US.In Egyptian environment MAS is delayed diagnoses,it is not diagnoses till the moment of birth, when the amniotic fluid appears greenish in color.Before birth, the fetal monitor may show a slow heart rate.Risk factors for this condition should be identified as early as possible.

Analysis shows that poor diagnose for blood disorder, which is mainly because of lack of facilities, is critical in Egypt. Accurate and early diagnosis of disease in Egypt may help in infant's treatments, decrease incubator length of stay.

Other factors were not included in Egyptian sample, due to their culture, like addicted mother.

5. CONCLUSION

A predictive models comparison have beencarried out, running on data from Egypt and United States, using different types of algorithms on the task of predicting incubator's length of stay. The models included several predictive algorithms i.e naïve bayes, support vector machine, logistic regression and linear regression. Differentperformance measurements have been used to evaluate the ability of the different models to accurately predict the length of stay. The results of the work have shown that in both environments, support vector machine outperforms other algorithms that were tried.

Comparing the results of running differentalgorithms on the two data sets shows convergence in predictive performance,except for logistic regression algorithm.Thissupports the use of the suggested data mining models with other data sources concerning preterm infants.

On other hand, the majority of factors listed in EGNN forms are common with factors in NHDS. However, some factors are reflecting the American environment; i.e. addicted mother.

Our research gives indication that early and accurate diagnoses of infant's diseases in Egyptian environment, i.e MAS and blood disorder, will speed up treatment and thus reduces length of stay.This will provides the optimal and best incubator usage for the specific infant case. The decision making for the length of stay has been improved and made more accurate.

The obtained results are expected to be useful for determining the LOS especially in the Egyptian environment.

6. ACKNOWLEDGMENTS

Authors thanks Prof.Dr. Mohammed El-Rewesha, director of EGNN Gharbia for his support and guidance, and Inter-University Consortium for Political and Social Research (ICPSR) for the use of the National Hospital Discharge Survey.

Also,Authors thanks Dr.Eng.Ahmed Marzook for his guidance in Oracle programming. Also, Hager Fady thanks Dr.Asmaa Sherief, assistant lecturer in forensic medicine and clinical toxicology department, Faculty of Medicine - TantaUniversity, for her help and cooperation in data collection**.**

REFERENCES

[1] R. Bellaz and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International Journal of Medical Informatics, vol. 77, no. 2, pp. 81-97, February 2008.

[2] H. Koh and G. Tan, "Data Mining Applications In Healthcare," J. Health. Info. Man, vol. 19, no. 2, pp. 64-72, 2005.

[3] WHO, "Newborn deaths decrease but account for higher share of global child deaths," 2011. [Online]. Available: http://www.who.int/mediacentre/news/releases/2011/newborn_deaths_20110830/en/index.html.

[4] J. Sandham, "Baby Incubation," 2008. [Online]. Available: http://www.ebme.co.uk.

[5] B. Zerinkow and K. Holtmannspötter, "Predicting Length-Of-Stay In Preterm Neonates," European Journal of Pediatrics, vol. 158, no. 1, 1999.

[6] S. Hintz, C. Bann, N. Ambalavanan, M. Cotten, A. Das and R. Higgins, "Predicting Time to Hospital Discharge for Extremely Preterm Infants," Journal of the American Academy of Pediatrics, vol. 125, pp. 146-154, 2010.

[7] P. Liu, L. Lei, J. Yin, W. Zhang, W. Naijun and E. El-Darzi, "Healthcare Data Mining: Prediction Inpatient Length of Stay," in 3rd International IEEE Conference on Intelligent Systems, Aveiro, 2006.

[8] G. Kraljevic and S. Gotovac, "Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services," Automatika, vol. 51, no. 3, pp. 275-283, 2010.

[9] EGNN, "Egyptian Neonatal Network," [Online]. Available: http://www.egynewborn.net.

[10] National Hospital Discharge Survey, 2002, Michigan: Inter-University Consortium for Political and Social Research, 2002.

[11] International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), Center of Disea Control, 2008.

[12] EGNN, "Dataset Manual," EGNN, Cairo, 2010.

[13] EGNN, "28 Day/Discharge Form," EGNN, Cairo, 2010.

[14] R. Haberstroh, Oracle® Data Mining Tutorial for Oracle Data Mining 11g Release 1, Oracle, 2008.

[15] M. Clinic, "Infant jaundice," Mayo Foundation for Medical Education and Research;, 2011. [Online]. Available: http://www.mayoclinic.com/health/infant-jaundice.

[16] J. Zurada and S. Lonial, "Comparison Of The Performance Of Several Data Mining Methods For Bad Debt Recovery In The Healthcare Industry," The Journal of Applied Business Research, vol. 21, no. 2, Spring 2005.