Progress Report for

ENGR 498 Global Design Project II

Analyzing a hospital’s data with machine learning techniques and predictions about the patient’s condition

Ahmet Cihad Taşpınar 214051453

Muhammed Ekrem Tecim 214101987

Muhammed Fatih Erdoğan 214000482

Supervisor: Mehmet Yasin Ulukuş

College of Engineering

İstanbul Şehir University

**Date: 18/05/2019**

Abstract:

In healthcare sector, there are enormous data volume collected in modern times thanks to clinical reports, doctor’s notes, body sensors, and also any kind of electronic (digital) mechanical devices. The data collected from biomedical sources are  getting larger and more complex every year. Thus, for better and more efficient medical services, data science and data analytics applications become more important. The analysis of health parameters and the estimation of future health conditions are in the informative phase yet. In this paper, we are focusing on hospitals’ intensive care patients data for predict their future conditions. For using recorded big data and machine learning techniques, we are going to try to clarify uncertainty and risk factors so that the hospital can use their resources efficiently.

Table of Contents

Introduction 3

Literature Review and Background 3

Body of Report 8

Evaluation 13

Conclusion 13

References 14

Introduction:

The data using health care management is considerable as big data due to its relevance on big data features which are called 5V- Volume, Velocity, Variety, Value, and Veracity. Currently, these data primarily used for monitoring patient's clinical situation for medical experts. The accumulated historical data also enables authorities (such as minister of health, or WHO) to statistical understanding as a result of descriptive statistic methods. When compared to these usage area of health data, it is possible to say that predictive and future oriented models are not common in launching system and decision making processes. Analyzing the structured and unstructured data generated from healthcare management systems with predictive statistical models and machine learning techniques for patient future conditions, death probabilities of patients and hospital capacity planning in some near future can lead to a fundamental change in medical area.

In our project, we analyzed an intensive care unit (ICU) which has a high proportion of cost in hospitals. With the collected data which has the features of patients in ICU such as oxygen level, patient blood pressure, blood electrolytes, heart rate, hemoglobin and hematocrit, and respiratory rate. With this collected data it has been created the derivative data. SOFA(Sequential Organ Failure Assessment) scores and Apache scores are calculated and located on derivative data. The derivative data consist of 69 columns and 16059 rows which are represented respectively features and number of patients. The analysis has been done by using this data. By following K fold Cross Validation technique data is separated in 10 folds and the algorithms are trained with 9 folds and success of algorithms are tested on rest fold. According to test result we can classify the patients into two groups considering their health parameters and we can predict whether they will alive or die.

In this report, firstly the table of literature review and background about machine learning techniques implementation in health sector will be given. Secondly, in the body of report, performance measures ROC/AUC and K-Fold Cross Validation will explained and then, the techniques are implemented in this project which are logistic regression and support vector machine will be clarified. In each techniques part the explanations implementations and codes with comments will included. In conclusion part, the success of two model will be compared and the test results will be interpreted. And the predictions about death situation of patients how can make efficiently for hospitals’ ICU unit operations and capacity planning.

Literature Review and Background:

The machine learning techniques has been implementing in health sector with the improvement of these study area. In literature, there are many studies focus on a specific health problem including in different clinics, model-interpreting failures, missing values… For our project we have examined the articles given below and try to understand them and get familiar the methods used and processes. The methods used in articles and explanation of them can be seen at the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Author & Years** | **Name of Article** | **Methods & Techniques** | **Explanation** |
| March 2017,  Teeradache Viangteeravat, PhD. Oguz Akbilgic, PhD, Robert Lowell Davis. | Analyzing Electronic Medical Records to Predict Risk of DIT (Death, Intubation, or Transfer to ICU) in Pediatric Respiratory Failure or Related Conditions | Multivariable logistic regression  K-mean clustering | They determine potential risk factors and their selection criteria such as Fraction of Inspired Oxygen (FiO2)† is risk factor  and Number of times FiO2 > 0.5 is selection criteria. After using multi logistic regression for these risk factors they cluster them and they reached the conclusion that race, respiratory rate, FiO2 rate and MCV  (Mean Corpuscular Volume of Blood Cell)ly associated with impending respiratory failure - defined in this study as intubation, transfer to the ICU, or death. Interestingly, the strongest predictor of impending DIT is MCV. |
| March 2018  Robert Steele,Nick Veith | Machine Learning-based Prediction of ICU Patient Mortality at Time of Admission | Logistic  SimpleLogistic  NaïveBayes  BayesNet    They built LazyKStar model.    WEKA software is used. | Using a large dataset  nearly 60,000 ICU encounters, they used machine learning algorithms to create predictive model and predict mortality of ICU patients. First they identified 7 attributes for inclusion in the model such as Admission Type(Urgent, Elective, Newborn or Emergency),Insurance,Marital Status,Religion,Language...  The models were evaluated using two different test options, a training set option and 10-fold cross validation (10-fold CV) according to the AUC metric. (produced an Area Under the Receiver Operating Characteristic Curve (AUC by machine learning models). LazyKStar algorithm has an AUC of 0.751 which has the highest when compared to other models. |
| 2016  Wiebke Sieben, Karen Schettlinger, Silvia Kuhls,Michael Imhof,Ursula Gather | Machine learning techniques in intensive care monitoring | Repeated Median filtering  Neyman-Pearson modified Random Forests | Filtering and Classification approaches for reducing intensive care units’ false alarms. Both  producers have ability reduce false alarm rate with high sensitivity. For practise, random Forests need more training data. Additionally, . robust repeated median filtering performs well for new patients. |
| January  2017  Prasan Kumar Sahoo, Suvendu Kumar Mohapatra, Shih-Lin Wu, | Analyzing Healthcare Big Data With Prediction for Future Health Condition | Intra-cluster Correlation Evaluation (IaCE),  Inter-Cluster Correlation Evaluation (IeCE),  Future Health Condition Prediction (FHCP). | By using probabilistic data collection method, correlation analysis of this collected data is performed.In conclusion, stochastic prediction method is built for forsee the future condition of patients based on current situation and their parametres. |
| June 2014,  Changwon Yoo, Luis Ramirez, Juan Liuzzi | Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine | Bayesian analysis  Continuous Models  Boolean Networks | This article focus on more relationship between data science and health industry. It mention about  Boolean Networks,Linear Regression,Logistic Regression and their potential usage of clinical, gene expression and bioinformatics area. Recently, the statistical methods that have been sued for study complex gene-gene, gene environment interactions. |
| January 2018  Jacob Fauber,Christian R. Shelton | Modeling “Presentness” of Electronic Health Record Data to Improve Patient State Estimation | The piecewise-constant conditional intensity model (PCIM) (Gunawardana et al., 2011)  Vs  Gaussian Process | After a PCIM model is trained on half the data (670 patient episodes), the monitor trajectories M are estimated using only the PCIM and the EHR(Electronic Health Records)  observations. Then they compare this average to a simple linear interpolation of the EHR observations, and a GP estimate trained in an analogous way.With EHR data, the problem is similar in structure but opposite in literal meaning; EHR data is absent almost everywhere. However, its presence is not at random, and this temporal information can be used to improve medical prediction. So the  “missing“ data problem in machine learning is tried to be eliminated. |
| August 2017  Hisashi Kurasawa, Akinori Fujino, and Katsuyoshi Hayashi | Predicting Patients’ Treatment Behavior by Medical Data Analysis Using Machine Learning Technique | Logistic regression model | The objective of article is develop a means of supporting decision making about treatment and medical prescription by doctors.And they focus on  designing a model to predict missed clinical appointment (MA) for identification of diabetes patients who need support. They use Tokyo Hospiltals’ historical data consist of of 16,026 clinical appointments. |
| June 2013  Peyman Rezaei Hachesu, PhD , Maryam Ahmadi, PhD , Somayyeh Alizadeh, PhD , Farahnaz Sadoughi, PhD | Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients | Classification Techniques with using three algorithm which are decision tree, support vector machines (SVM), and artificial neural network (ANN). | Length of stay (LOS) of patients in a hospital is important for in terms of quality of service and customer satisfaction.In this article, it is try to be estimated that  LOS of heart patients by data mining techniques.The patient records of 4,948 patients who had suffered CAD ( coronary artery disease ) is used in this article.In conclusion, all three algorithm predict LOS in different level of accuracy.  Finding shows SVM was the best fit.LOS increases in patients who has h lung or respiratory disorders and high blood pressure. |
| 2018  Murali Ravuri,Anitha Kannan, Geoffrey J. Tso, Xavier Amatriain | Learning from the experts: From expert systems to machine-learned diagnosis models | Multiclass logistic regression,  Deep neural network model,  Experiments and simulations | Expert diagnostic support systems are one of  intense study area of AI. The practical applications are limited due to lack of extensibility. Machine-learned models for medical diagnosis become more important after these models can learn and generalize patterns found in very large data like ERHs.In this paper, writers try to use EHR dataset for an expert knowledge base diagnosis models. It is a  complicated study. |
| 2018  Selin Merdan,Khurshid Ghani ,Brian Denton | Integrating Machine Learning and Optimization Methods for Imaging of Patients with Prostate Cancer | Decision tree for designing coordinated imaging protocols,  The Robust Model | In this paper, writers combined optimization models into a robust optimization framework and  predictive models from machine learning for designing image guidelines. They incorporated the perspectives of patients and physicians at the population level, and proposed models for sequential testing where the outcome of one imaging test informs the decision about the follow-up test. In addition to the optimization models, they proposed clinically motivated approximation algorithms. |
| 2018  Muhammad A Ahmad, Carly Eckert, Greg McKelvey, Kiyana Zolfagar, Anam Zahid, Ankur Teredesai | Death versus Data Science: Predicting End of Life | Decision Trees (Quinlan 1986),  Bayesian Rule Lists (Yang, Rudin, and Seltzer 2016),  Regression Trees (De’ath and Fabricius 2000) | In this article, it is tried to predict t risk of mortality for patients from two large hospital systems in the Pacific Northwest. Using medical claims and electronic medical records (EMR) data. |
| 2013  G. James  D. Witten  T. Hastie  R. Tibshirani | An Introduction to Statistical Learning | Supoort Vector Machines, Logistic Regression, R Studio, Resampling Methods | In this book, majority of the statistical learning techniques and their applications in R studio are tried to explain. |

Body of the Report:

Machine learning algorithms are separated to three main groups which are supervised, unsupervised and reinforcement learning. These separations are based on online vs. batch learning or instance based vs. model based algorithms. In supervised learnings training is labeled which means for each instance in training data right answers called labels are given. However, in supervised learning training data is unlabeled so the system tries to learn without teacher. In reinforcement learning, the main goal is to develop a system that improves its performance based on interactions with the environment. In short, supervised learning have labeled data and predict outcomes, unsupervised learning don’t have labels and find the hidden structure in data, and reinforcement learning is a kind of decision process and learns series of actions.

We applied support vector machine and logistic regression techniques to our medical data which are sub-techniques of supervised machine learning. We also took a glance at the resampling methods and used k-fold cross validation which is an essential resampling method. Validation will provide us to use all of the data as test set and more than one evaluation, which is a better performance measurement. Moreover, one of the most important reason why we will use validation is that, we should find the best parameters to reach the best accuracy and ROC curves, which is done by a validation set.

AUC/ROC:

        In Machine Learning, performance measurement is one of the most important task for understanding the success of learning . So, when a classification problem appears the performance of classifying measures by AUC - ROC Curve. Also the need for checking or visualizing the performance of the multi - class classification problem is resolved by AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. ROC is a probability curve and AUC represents degree or measure of separability. It basically explains the degree of models’ distinguish capabilities between classes.



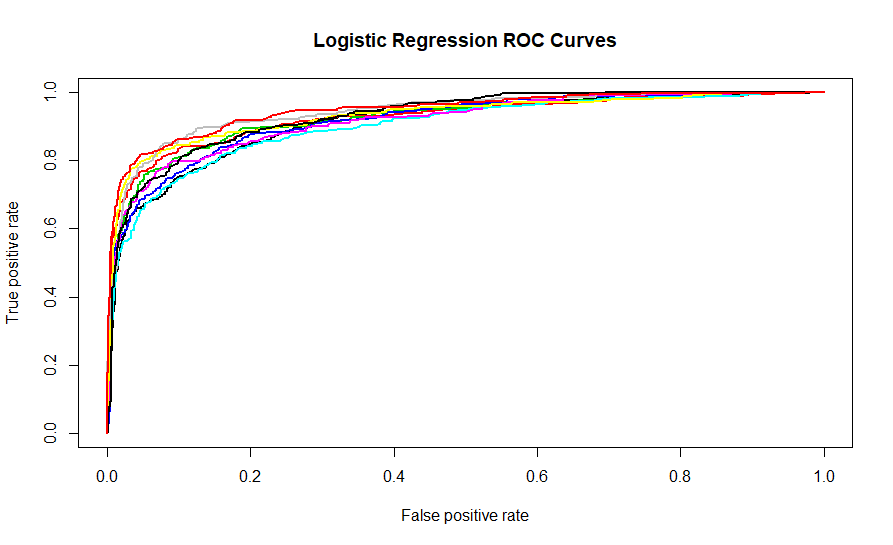
For creating ROC, false positive rate and true positive rate is need to be calculated. For this calculation confusion matrix has to be created. Basically confusion matrix shows predicted values and actual values. By using this matrix Specificity, Accuracy, Recall and Precision can be calculated.

The model has the higher AUC score is the better predicting tools for classifying. By analogy, Higher the AUC, better the model is at distinguishing between patients have similar heath parameters whether they die or alive.

Logistic Regression:

Logistic regression predicts the probability of a result that can have only two values. The prediction is based on the use of one or several predictors (numerical and categorical). Linear regression is not appropriate for the values that can be expressed in binary system as yes / no, exist / not. Because, it can estimate the value outside the range of 0 and 1. Logistic regression produces a limited logistic curve with values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is created using the natural logarithm of the probabilities of the target variable instead of the probability.

Logistic Regression Model:



Mean AUC=0.92364953

The result of 10-Fold Cross Validation and ROC curve of each fold is above. The mean AUC of 10 folds is 0.92364953 which represent the success of our model. Logistic Regression Model learns from 14453 patients’ health parameter and it has 0.92364953 success of predicting new patients’ death case. That means this model predict the patients who are going to die as die and it also predict the patients who are going to live as live at the rate of 0.92.

Model Summary:

In the model summary given below, the stars on the right in the table are the shows how significant is the factor by its p-value. Factors that marked with three stars are the most significant of all.

Call:

glm(formula = trainData$death ~ ICU\_Type + Dest\_Level\_of\_Care +

age + sex + LOS + Initial\_SOFA\_Liver + Initial\_SOFA\_Coagulation +

Initial\_SOFA\_Nerv + Initial\_SOFA\_Renal + Initial\_SOFA\_Respiratory +

Initial\_SOFA\_Cardio + Discharge\_SOFA\_Liver + Discharge\_SOFA\_Coagulation +

Discharge\_SOFA\_Nerv + Discharge\_SOFA\_Renal + Discharge\_SOFA\_Respiratory +

Discharge\_SOFA\_Cardio + Max\_SOFA\_Liver + Max\_SOFA\_Coagulation +

Max\_SOFA\_Nerv + Max\_SOFA\_Renal + Max\_SOFA\_Respiratory + Max\_SOFA\_Cardio +

Aver\_SOFA\_Liver + Aver\_SOFA\_Coagulation + Aver\_SOFA\_Nerv +

Aver\_SOFA\_Renal + Aver\_SOFA\_Respiratory + Aver\_SOFA\_Cardio +

Var\_SOFA\_Liver + Var\_SOFA\_Coagulation + var\_SOFA\_Nerv + Var\_SOFA\_Renal +

Var\_SOFA\_Respiratory + Var\_SOFA\_Cardio + AdmitApache + Charlson\_index +

cvc\_status + Type + disch\_night + weekend + previous\_ICU\_stays +

SIRS\_48\_hour + MV\_24\_hour, family = binomial, data = trainData)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.5293 -0.3839 -0.2260 -0.1112 3.2435

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.9503103 0.3276153 -15.110 < 2e-16 \*\*\*

ICU\_TypeCardiac-T -0.4837984 0.1437172 -3.366 0.000762 \*\*\*

ICU\_TypeMedical 0.2993019 0.1177981 2.541 0.011060 \*

ICU\_TypeNeuro 0.2011943 0.1119056 1.798 0.072194 .

ICU\_TypeSurgical -0.2119033 0.1206785 -1.756 0.079100 .

ICU\_TypeTransplant -0.3505639 0.1330065 -2.636 0.008397 \*\*

ICU\_TypeTrauma -0.0452494 0.2048974 -0.221 0.825217

Dest\_Level\_of\_CareRegular -1.2649876 0.1053890 -12.003 < 2e-16 \*\*\*

Dest\_Level\_of\_CareSDU -1.2466469 0.0880607 -14.157 < 2e-16 \*\*\*

age 0.0313411 0.0020809 15.061 < 2e-16 \*\*\*

sexM -0.0443277 0.0640572 -0.692 0.488935

LOS 0.0001641 0.0001426 1.151 0.249936

Initial\_SOFA\_Liver -0.0568560 0.1129713 -0.503 0.614768

Initial\_SOFA\_Coagulation -0.1085208 0.0711654 -1.525 0.127281

Initial\_SOFA\_Nerv -0.1186000 0.0354069 -3.350 0.000809 \*\*\*

Initial\_SOFA\_Renal 0.0081287 0.0683512 0.119 0.905335

Initial\_SOFA\_Respiratory 0.1097611 0.0631622 1.738 0.082252 .

Initial\_SOFA\_Cardio 0.0698065 0.0501584 1.392 0.164007

Discharge\_SOFA\_Liver 0.7218522 0.1403389 5.144 2.69e-07 \*\*\*

Discharge\_SOFA\_Coagulation 0.1794979 0.0820221 2.188 0.028640 \*

Discharge\_SOFA\_Nerv 0.8066781 0.0462945 17.425 < 2e-16 \*\*\*

Discharge\_SOFA\_Renal 0.2523842 0.0882082 2.861 0.004220 \*\*

Discharge\_SOFA\_Respiratory 0.3539121 0.0822376 4.304 1.68e-05 \*\*\*

Discharge\_SOFA\_Cardio 0.4589591 0.0594126 7.725 1.12e-14 \*\*\*

Max\_SOFA\_Liver 0.0293451 0.1732887 0.169 0.865527

Max\_SOFA\_Coagulation -0.0038859 0.1000693 -0.039 0.969024

Max\_SOFA\_Nerv -0.1707806 0.0600818 -2.842 0.004477 \*\*

Max\_SOFA\_Renal 0.1106276 0.1120985 0.987 0.323702

Max\_SOFA\_Respiratory -0.0993369 0.0788597 -1.260 0.207790

Max\_SOFA\_Cardio -0.0129615 0.0643449 -0.201 0.840356

Aver\_SOFA\_Liver -0.3021660 0.2162924 -1.397 0.162406

Aver\_SOFA\_Coagulation 0.1754476 0.1460918 1.201 0.229774

Aver\_SOFA\_Nerv 0.0524296 0.0744837 0.704 0.481491

Aver\_SOFA\_Renal -0.2019512 0.1360860 -1.484 0.137810

Aver\_SOFA\_Respiratory -0.3676061 0.1262575 -2.912 0.003596 \*\*

Aver\_SOFA\_Cardio -0.0875762 0.0954142 -0.918 0.358696

Var\_SOFA\_Liver -0.5168411 0.3365713 -1.536 0.124635

Var\_SOFA\_Coagulation -0.3257680 0.2030513 -1.604 0.108634

var\_SOFA\_Nerv -0.1787060 0.0703544 -2.540 0.011083 \*

Var\_SOFA\_Renal 0.3248522 0.2150508 1.511 0.130895

Var\_SOFA\_Respiratory 0.7784363 0.1791906 4.344 1.40e-05 \*\*\*

Var\_SOFA\_Cardio -0.0854287 0.0897752 -0.952 0.341307

AdmitApache 0.0116285 0.0027452 4.236 2.28e-05 \*\*\*

Charlson\_index 0.0584846 0.0994881 0.588 0.556629

cvc\_status -0.0361999 0.0731401 -0.495 0.620643

TypeSURG -0.2072435 0.0673895 -3.075 0.002103 \*\*

disch\_night 0.4167299 0.0686232 6.073 1.26e-09 \*\*\*

weekend 0.2656183 0.0707195 3.756 0.000173 \*\*\*

previous\_ICU\_stays -0.0228938 0.0670866 -0.341 0.732910

SIRS\_48\_hour 0.4849426 0.0643360 7.538 4.79e-14 \*\*\*

MV\_24\_hour 0.0356902 0.0830782 0.430 0.667489

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

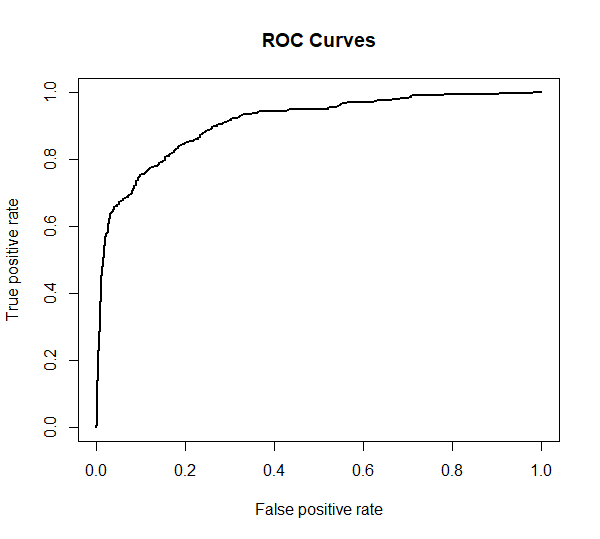
Null deviance: 14192.7 on 14452 degrees of freedom

Residual deviance: 7138.4 on 14402 degrees of freedom

AIC: 7240.4

Number of Fisher Scoring iterations: 6

ROC Curve for Each Fold:



[[1]]

Accuracy: 0.896638

AUC Result: 0.9087741

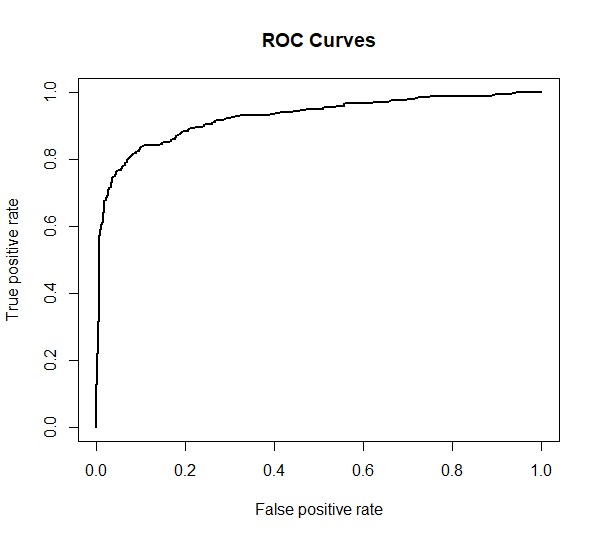
Confusion Matrix:

y\_pred

y\_true 0 1

0 1228 49

1 117 212



[[2]]

Accuracy: 0.918431

AUC Result: 0.9245504

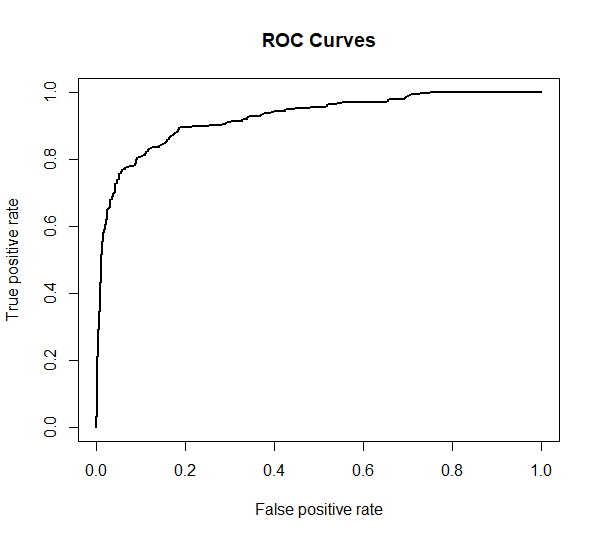
Confusion Matrix:

y\_pred

y\_true 0 1

0 1256 36

1 95 219



[[3]]

Accuracy: 0.912204

AUC Result:0.9233916

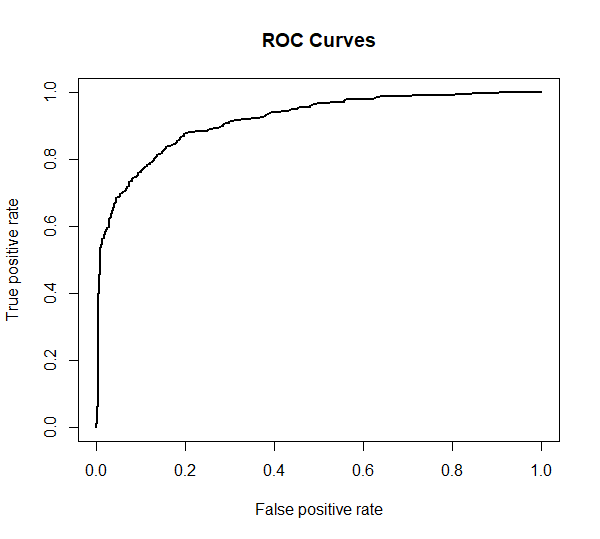
Confusion Matrix:

y\_pred

y\_true 0 1

0 1264 40

1 101 201



[[4]]

Accuracy: 0.900374

AUC Result: 0.9161595

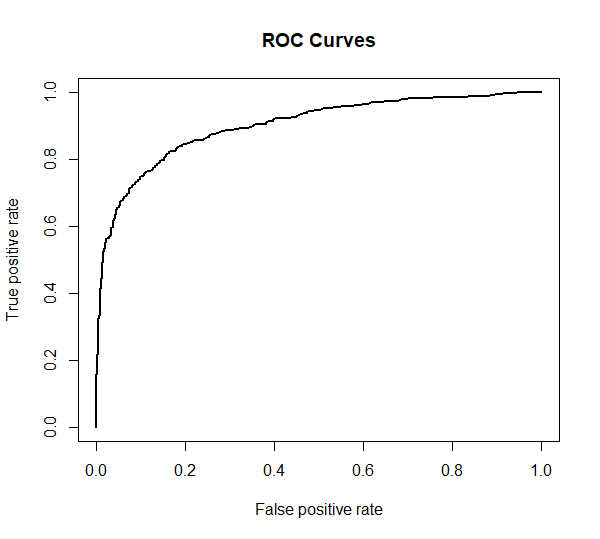
Confusion Matrix:

y\_pred

y\_true 0 1

0 1239 52

1 108 207



[[5]]

Accuracy: 0.888543

AUC Result: 0.8999539

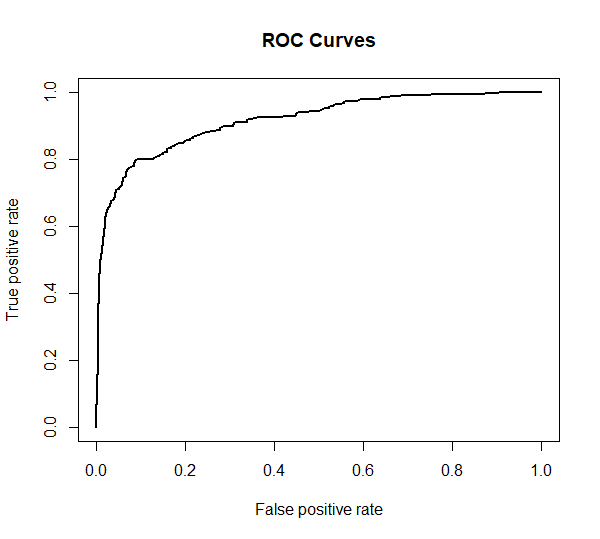
Confusion Matrix:

y\_pred

y\_true 0 1

0 1221 48

1 131 206



[[6]]

Accuracy: 0.911582

AUC Result: 0.9150761

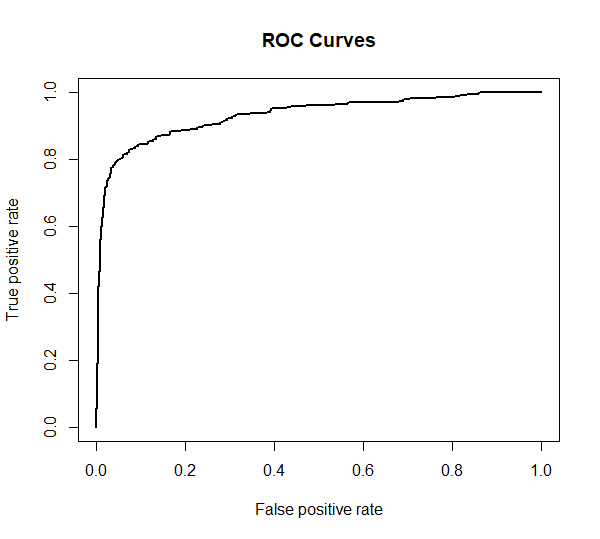
Confusion Matrix:

y\_pred

y\_true 0 1

0 1263 35

1 106 201



[[7]]

Accuracy: 0.929016

AUC Result: 0.9305641

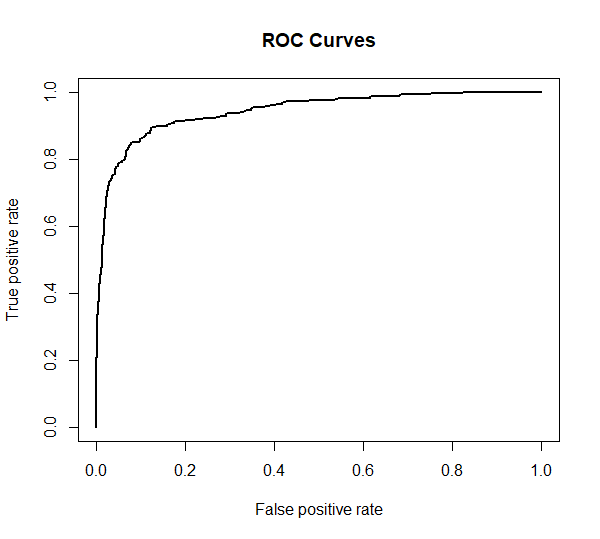
Confusion Matrix:

y\_pred

y\_true 0 1

0 1279 27

1 87 213



[[8]]

Accuracy: 0.927771

AUC Result: 0.9421554

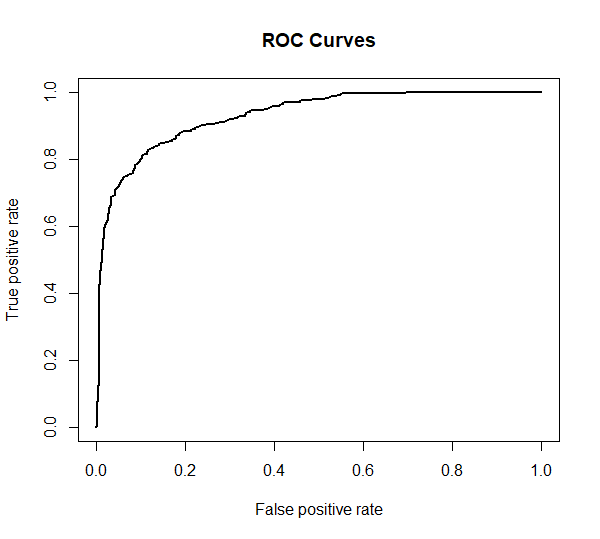
Confusion Matrix:

y\_pred

y\_true 0 1

0 1290 35

1 81 200



[[9]]

Accuracy: 0.909091

AUC Result: 0.9296335

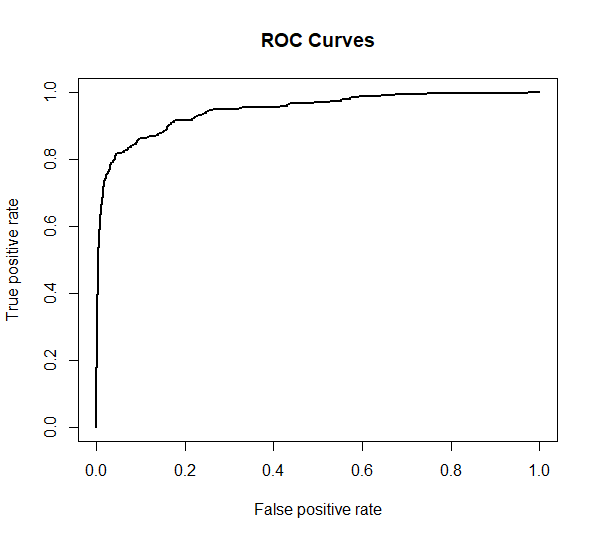
Confusion Matrix:

y\_pred

y\_true 0 1

0 1257 40

1 106 203



[[10]]

Accuracy: 0.935243

AUC Result: 0.9462367

Confusion Matrix:

y\_pred

y\_true 0 1

0 1296 21

1 83 206

When we examine the 10 folds seperately, we see that the best result comes in the 10th fold with 0.935243 accuracy and 0.9462367 AUC while, the worst is the 5th fold with accuracy 0.888543 and 0.8999539 AUC. Actually, even the worst case’s accuracy and AUC is fine however, model is made to deal with human life conditions so, the objective of accuracy or AUC should be as close as possible to maximum.

R Code for Logistic Regression Model:

library(e1071)

library(ISLR)

library(caret)

library (ROCR )

library (cutpointr)

library("MLmetrics")

#data is read

read\_data<-read.csv("ICUPatients.csv")

#str(read\_data)

#summary(read\_data)

#main features are assigned

Features<-c("ICU\_Type", "Dest\_Level\_of\_Care", "age", "sex", "LOS",

"Initial\_SOFA\_Liver", "Initial\_SOFA\_Coagulation", "Initial\_SOFA\_Nerv","Initial\_SOFA\_Renal","Initial\_SOFA\_Respiratory","Initial\_SOFA\_Cardio",

"Discharge\_SOFA\_Liver","Discharge\_SOFA\_Coagulation","Discharge\_SOFA\_Nerv","Discharge\_SOFA\_Renal","Discharge\_SOFA\_Respiratory","Discharge\_SOFA\_Cardio",

"Max\_SOFA\_Liver","Max\_SOFA\_Coagulation","Max\_SOFA\_Nerv","Max\_SOFA\_Renal","Max\_SOFA\_Respiratory","Max\_SOFA\_Cardio",

"Aver\_SOFA\_Liver","Aver\_SOFA\_Coagulation","Aver\_SOFA\_Nerv","Aver\_SOFA\_Renal","Aver\_SOFA\_Respiratory","Aver\_SOFA\_Cardio",

"Var\_SOFA\_Liver","Var\_SOFA\_Coagulation","var\_SOFA\_Nerv","Var\_SOFA\_Renal","Var\_SOFA\_Respiratory","Var\_SOFA\_Cardio",

"AdmitApache","Charlson\_index","cvc\_status","Type","disch\_night","weekend","previous\_ICU\_stays","SIRS\_48\_hour","MV\_24\_hour"

,"death" )

#main data frame is created

main\_df<-read\_data[Features]

#summary(main\_df)

#str(main\_df)

#View(main\_df)

#Data Slicing

set.seed(1) #for creating rondom numbers assigned a seed number

#K-fold cross-validation

folds <- cut(seq(1,nrow(main\_df)),breaks=10,labels=FALSE)

#Perform 10 fold cross validation

for(i in 1:10){

#Segement your data by fold using the which() function

testIndexes <- which(folds==10,arr.ind=TRUE)

testData <- main\_df[testIndexes, ]

trainData <- main\_df[-testIndexes, ]

#Use the test and train data partitions however you desire...

#Model Building

glm.fit =glm (trainData$death~ICU\_Type+Dest\_Level\_of\_Care+age+sex+LOS+

Initial\_SOFA\_Liver+Initial\_SOFA\_Coagulation+Initial\_SOFA\_Nerv+Initial\_SOFA\_Renal+Initial\_SOFA\_Respiratory+Initial\_SOFA\_Cardio+

Discharge\_SOFA\_Liver+Discharge\_SOFA\_Coagulation+Discharge\_SOFA\_Nerv+Discharge\_SOFA\_Renal+Discharge\_SOFA\_Respiratory+Discharge\_SOFA\_Cardio+

Max\_SOFA\_Liver+Max\_SOFA\_Coagulation+Max\_SOFA\_Nerv+Max\_SOFA\_Renal+Max\_SOFA\_Respiratory+Max\_SOFA\_Cardio+

Aver\_SOFA\_Liver+Aver\_SOFA\_Coagulation+Aver\_SOFA\_Nerv+Aver\_SOFA\_Renal+Aver\_SOFA\_Respiratory+Aver\_SOFA\_Cardio+

Var\_SOFA\_Liver+Var\_SOFA\_Coagulation+var\_SOFA\_Nerv+Var\_SOFA\_Renal+Var\_SOFA\_Respiratory+Var\_SOFA\_Cardio+

AdmitApache+Charlson\_index+cvc\_status+Type+disch\_night+weekend+previous\_ICU\_stays+SIRS\_48\_hour+MV\_24\_hour, data=trainData, family = binomial )

#Fitting the model

fitted = predict(glm.fit,testData,type="response")

summary(glm.fit)

rocplot <- function (fitted,truth){

predob = prediction (as.numeric(fitted),as.numeric(truth))

perf = performance (predob, "tpr", "fpr")

plot(perf, plotCI.lwd = 2, lwd = 2,col=as.list(1:10),spread.estimate = "boxplot", main="ROC Curves")

AUC = performance(predob,"auc")

print(AUC@y.values)

}

#Plotting the ROC Curves

rocplot(fitted,testData$death)

#Printing Confusion Matrixs

pred <- ifelse(fitted < 0.5, 0, 1)

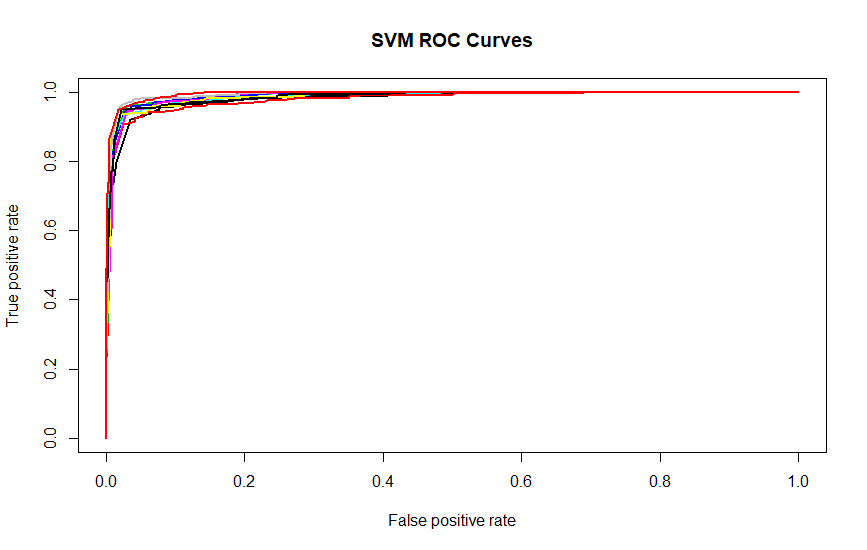
conmatrix=ConfusionMatrix(y\_pred = pred, y\_true = testData$death)

print(conmatrix)

}

Support Vector Machines:

SVM is a powerful supervised learning algorithm that can be used in both regression and classification solutions. For classification, it is possible to separate two groups by drawing a border between two groups in a plane. The place where this limit will be drawn should be the most distant place for the members of both groups. Here SVM determines how this limit is drawn. In order to carry out this process, two boundary lines are drawn close and parallel to each other and these boundary lines are brought together and a common boundary line is produced. We will use SVM to separate our patients to 2 or 3 groups which will tell us the conditions of them. Therefore, there should be a patient score to predict patients groups.



Mean AUC Result: 0.98750148

Call:

svm(formula = y ~ ., data = trainData, kernel = "radial", gamma = 1, cost = 1)

Parameters:

SVM-Type: eps-regression

SVM-Kernel: radial

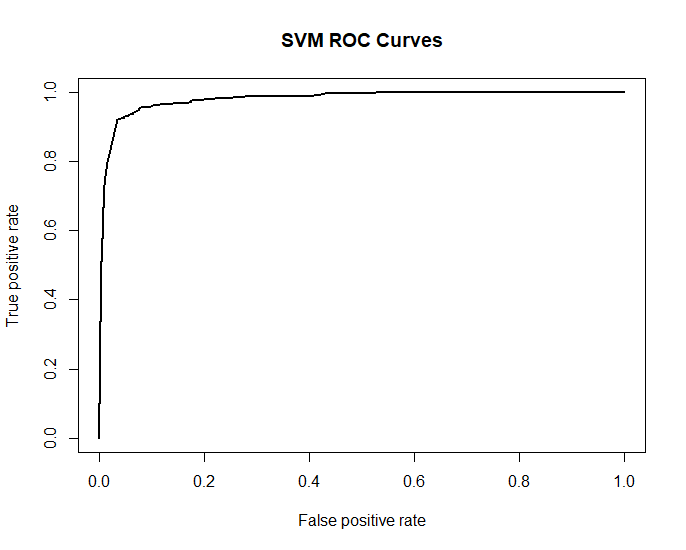
cost: 1

gamma: 1

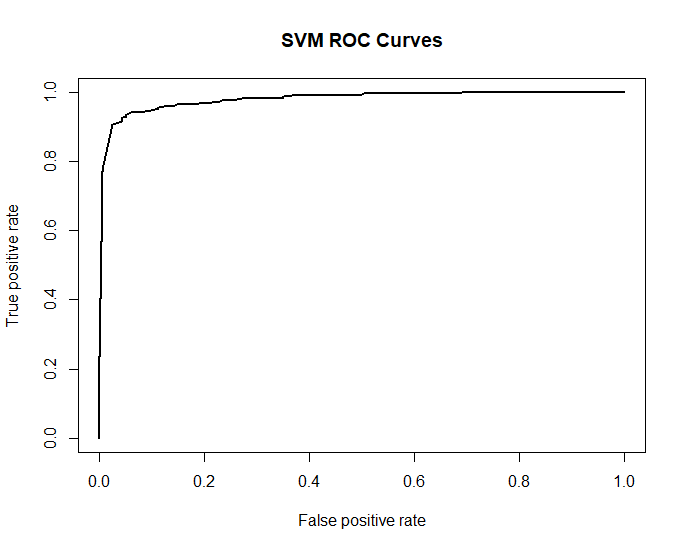
epsilon: 0.1

Number of Support Vectors: 14433

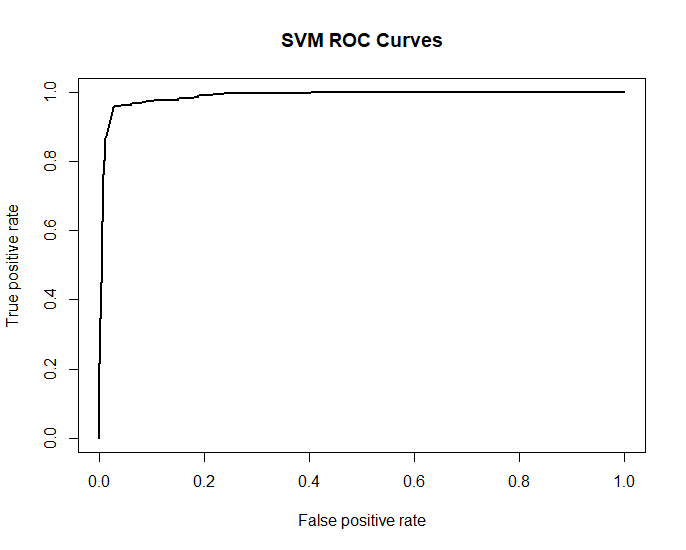
ROC Curve for Each Fold:



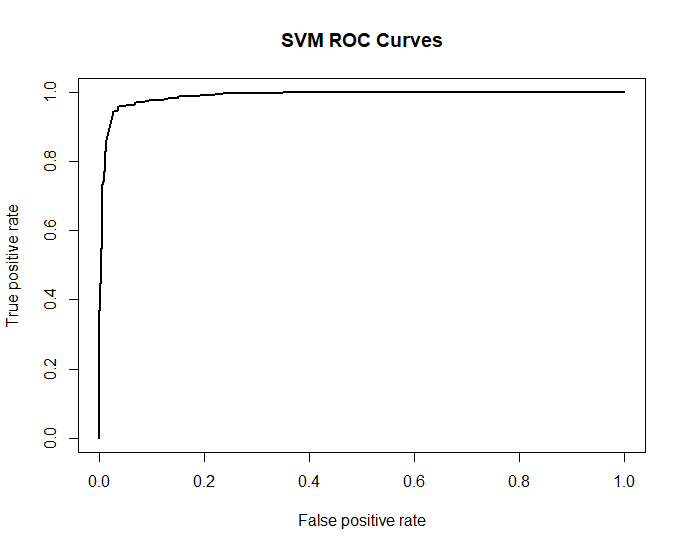
AUC Result: 0.9809049



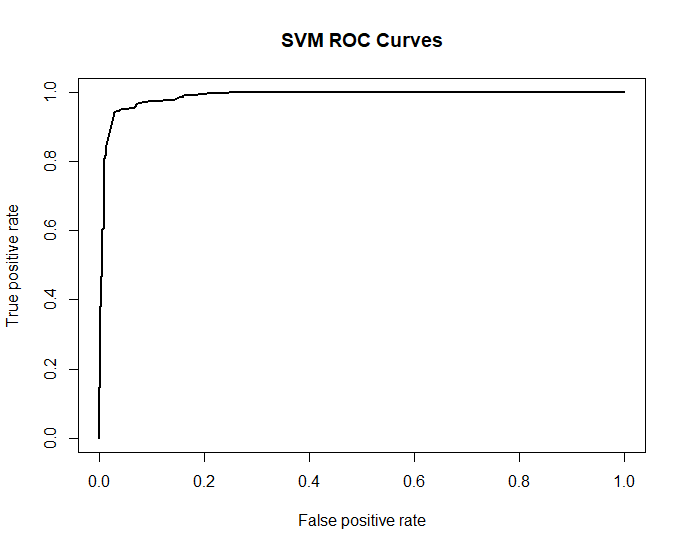
AUC Result: 0.9797862



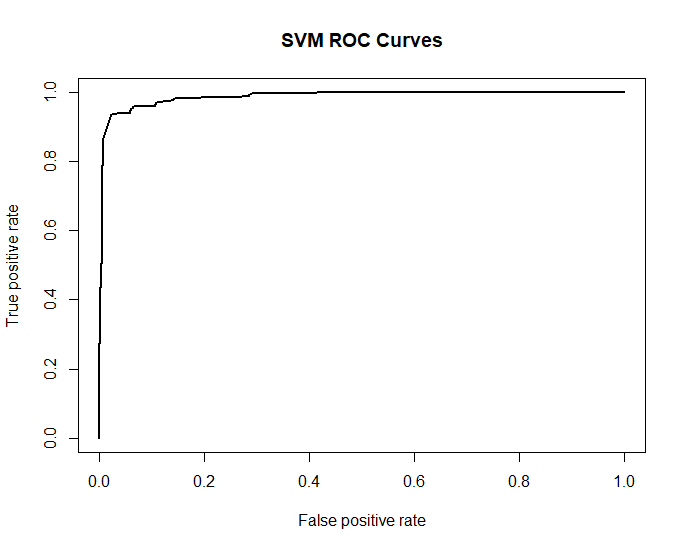
AUC Result: 0.9885363



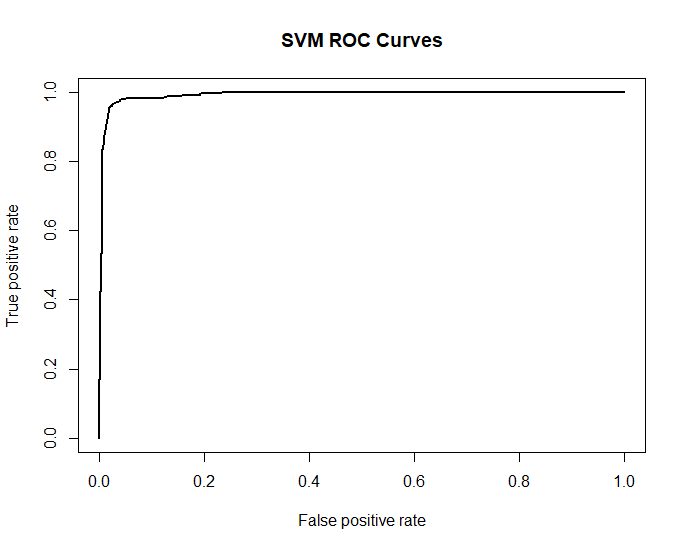
AUC Result:  0.9892307



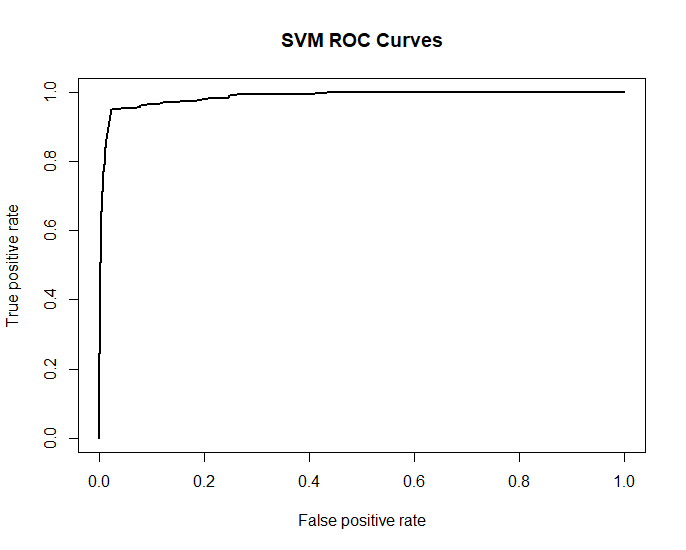
AUC Result: 0.9875881



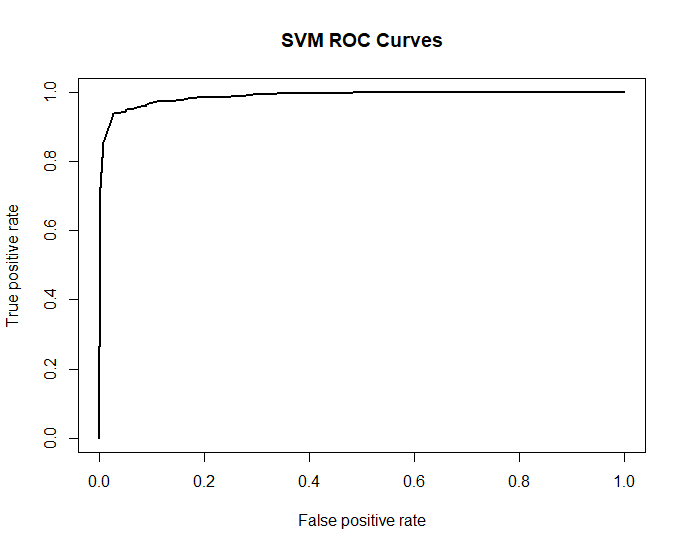
AUC Result: 0.9879331



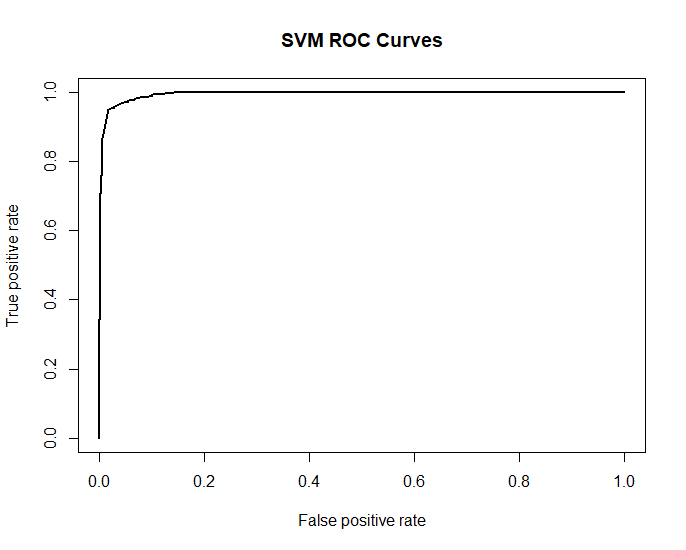
AUC Result: 0.9871937



AUC Result:  0.9924649



AUC Result: 0.9867157



AUC Result: 0.9946612

R Code for SVM Model:

library(e1071)

library(ISLR)

library(caret)

library (ROCR )

library(plyr)

library(MLmetrics)

#Data is read

read\_data<-read.csv("ICUPatients.csv")

#str(read\_data)

#summary(read\_data)

#Main features are assigned

Features<-c("ICU\_Type", "Dest\_Level\_of\_Care", "age", "sex", "LOS",

"Initial\_SOFA\_Liver", "Initial\_SOFA\_Coagulation", "Initial\_SOFA\_Nerv","Initial\_SOFA\_Renal","Initial\_SOFA\_Respiratory","Initial\_SOFA\_Cardio",

"Discharge\_SOFA\_Liver","Discharge\_SOFA\_Coagulation","Discharge\_SOFA\_Nerv","Discharge\_SOFA\_Renal","Discharge\_SOFA\_Respiratory","Discharge\_SOFA\_Cardio",

"Max\_SOFA\_Liver","Max\_SOFA\_Coagulation","Max\_SOFA\_Nerv","Max\_SOFA\_Renal","Max\_SOFA\_Respiratory","Max\_SOFA\_Cardio",

"Aver\_SOFA\_Liver","Aver\_SOFA\_Coagulation","Aver\_SOFA\_Nerv","Aver\_SOFA\_Renal","Aver\_SOFA\_Respiratory","Aver\_SOFA\_Cardio",

"Var\_SOFA\_Liver","Var\_SOFA\_Coagulation","var\_SOFA\_Nerv","Var\_SOFA\_Renal","Var\_SOFA\_Respiratory","Var\_SOFA\_Cardio",

"AdmitApache","Charlson\_index","cvc\_status","Type","disch\_night","weekend","previous\_ICU\_stays","SIRS\_48\_hour","MV\_24\_hour"

,"death" )

#View(Features)

#View(read\_data[Features])

#summary(read\_data[Features])

#for checking missing values

#anyNA(read\_data[Features])

#data cleaning and preapre features data = main\_df

#Main data frame is created

main\_df<-read\_data[Features]

#summary(main\_df)

#str(main\_df)

#View(main\_df)

#Dataslicing,K-Fold Cross Validation and Roc Curves

set.seed(1)

#for creating rondom numbers assigned a seed number

folds <- cut(seq(1,nrow(main\_df)),breaks=10,labels=FALSE)

#Perform 10 fold cross validation

#This method is used to make our work replicable

for(i in 1:10){

#Segement your data by fold using the which() function

testIndexes <- which(folds==1,arr.ind=TRUE)

testData <- main\_df[testIndexes, ]

#View(testData)

#Test data is created

trainData <- main\_df[-testIndexes, ]

#View(testData)

#Train data is created

y = trainData$death

#Considering the death results of patient, we try to find a separator.

svmfit =svm (y~. ,data= trainData, kernel ="radial", gamma =1,

cost =1)

summary(svmfit)

#Primary model is built and its properties are summarized.

tuned\_parameters <- tune.svm(death~., data = trainData, gamma = 10^(-3:1), cost = 10^(-3:3))

summary(tuned\_parameters )

#Tune.svm is a generic function tunes hyperparameters of statistical methods using a grid search over supplied parameter ranges.

#At the given parameter range tune.svm finds best parameters and performances.

fitted = predict(svmfit,testData, type="response")

#Predicting the model success into test data.

rocplot <- function (fitted,truth){

predob = prediction (as.numeric(fitted), as.numeric(truth))

perf = performance (predob, "tpr", "fpr")

plot(perf,add=TRUE, main="SVM ROC Curves",plotCI.lwd = 2, lwd = 2,diag=TRUE)

AUC = performance(predob,"auc")

print(AUC@y.values)

#Area under curve values are printed.

}

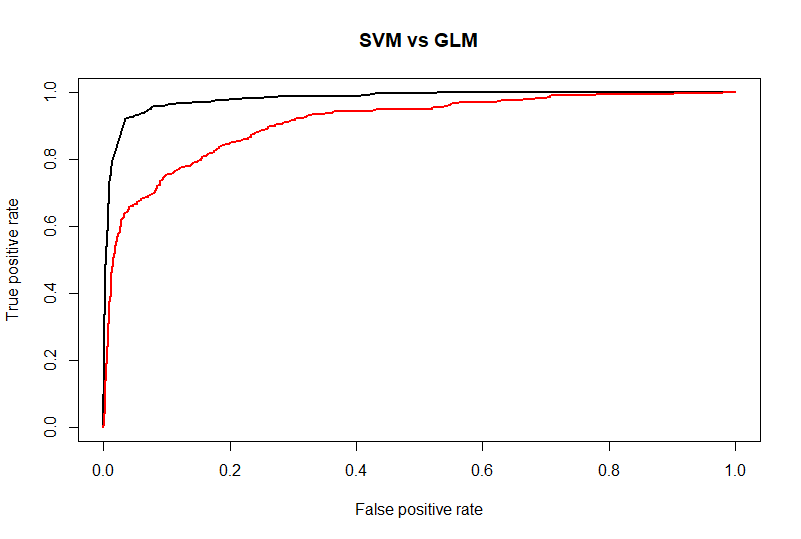
rocplot(fitted,testData$death)

#ROC Curves are printed.

}

Conclusion

After searching and implementing some machine learning algorithms for predicting mortality of ICU patients we can conclude that predicting death completely and without derogating is extremely hard. But with developing technology and data science applications on big data enable people to analyze various medical parameter and it is successful for creating a general impression about patient situation. Even there are some existing algorithms can calculate death rate and predict patients’ mortality, with building two different models with machine learning methods(Support Vector Machine and Logistic Regression) we estimated the mortality of incoming patients. The performance of estimating patients’ death status is higher in Support Vector Machine Model compared to the Logistic Regression Model.



In conclusion part, the success of two model will be compared and the test results will be interpreted. And the predictions about death situation of patients how can make efficiently for hospitals’ ICU unit operations and capacity planning.

References:

·      Viangteeravat, T., Akbilgic, O., & Davis, R. L. (2017). Analyzing Electronic Medical Records to Predict Risk of DIT (Death, Intubation, or Transfer to ICU) in Pediatric Respiratory Failure or Related Conditions. *AMIA Summits on Translational Science Proceedings*, *2017*, 287.

·     Yoo, C., Ramirez, L., & Liuzzi, J. (2014). Big data analysis using modern statistical and machine learning methods in medicine. *International neurourology journal*, *18*(2), 50.

·      Sahoo, P. K., Mohapatra, S. K., & Wu, S. L. (2016). Analyzing Healthcare Big Data With Prediction for Future Health Condition. *IEEE Access*, *4*, 9786-9799.

·      Ahmad, M. A., Eckert, C., McKelvey, G., Zolfaghar, K., Zahid, A., & Teredesai, A. (2018). Death vs. Data Science: Predicting End of Life. In *AAAI*.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.