**PREDICTIVE ANALYSIS ON SURVIVAL OF ICU PATIENTS**

Adithya S Sukhesh K N Mahesh P

II MSc Statistics II MSc Statistics II MSc Statistics

[adithyasagri@gmail.com](mailto:adithyasagri@gmail.com), [sukheshkn241199@gmail.com](mailto:sukheshkn241199@gmail.com) [mmp9021@gmail.com](mailto:mmp9021@gmail.com)

Department of statistics

Mangalore University

Mangalagangotri, Karnataka -574199

**ABSTRACT:**

Identifying the factors affecting the survival of ICU patients and accurate prediction of survival probability of patient admitted to ICU plays important role which helps patient family to decide whether to go for further treatment or not. The biggest challenge for researcher in the field of medical science is to identify and model the direct and indirect effect between variables simultaneously and effectively. An empirical study is carried out to show the effectiveness of path modelling to handle complex nature among variables to predict survival of ICU patients. The Predictive performance of path model is compared with Logistic regression model, Decision tree classification, Random forest, XGBoost model, K-Nearest Neighbor classification, Support vector machine. The result shows that path modelling outperforms all other classification models and can be used to predict survival of ICU patient accurately.

**Keywords:** ICU, medical field, Logistic regression, Path modelling, Decision tree, Random forest, XGBoost model, K-Nearest Neighbor classification, Support vector classifier

1. **INTRODUCTION**

An intensive care refers to the specialized treatment given to patients who are extremely unwell and require critical medical care. An intensive care unit (ICU) provides the critical care and life support for acutely ill and injured patients from well trained and experienced doctors, nurses’ .ICU patients should get treatment in time as their mortality rate will be high as compared to other patients in hospital. The aim of the ICU is further prevention of morbidity and mortality of patients. But, in fact the patients, who really require a special care are not getting proper facilities of ICU. Rather than the needy, the patients who have very low mortality rate are getting the facilities. This might be care taker’s lack of proper knowledge about ICU.

Treating patients in ICU is expensive which includes total cost of all resources- clinical and administrative personnel, , devices, drugs and other supplies, and equipment used during a patient’s full cycle of care for a specific medical condition. Most of the common people cannot afford such a huge amount without knowing whether the patient survives or not. Therefore, the proper study and correct detection of the chance of survival of ICU patients using an efficient classification model will reduce the burden of doctors as well as avoid great loss.

Compared to Traditional Statistical methods, Machine learning methods are extensively applied to various health science datasets to identify the hidden pattern and a predictive classifier can be constructed for future decision making. Many studies has been carried out to show the efficiency of machine learning models. A study on Comparing Machine Learning algorithms for predicting ICU admission and mortality in COVID – 19 carried by Sonu Subudhi (2020) shows that ensemble models performs better than Logistic Regression, KNN, Linear Discriminant analysis, MLP Classifier for predicting the survival rate of ICU patients. Sujeong et al (2021) suggested a Unplanned extubation (UE) prediction model using machine learning algorithms, which included, logistic regression (LR), Random forest (RF), and support vector machine (SVM).

Ryoung et al (2021) applied the Machine learning models such as Extreme Gradient Boosting (XGBoost), Random forest (RF), Deep neural network (DNN), and logistic regression (LR) for the prediction of Intensive Care Unit Delirium. Siyi Yuan et al (2021) advocated XGBoost, Support Vector Machine (SVM), Random Forest (RF), ExtraTrees (ET) models to predict Candidaemia in ICU Patients with New-Onset Systematic Inflammatory Response Syndrome and conclude that XGBoost performs better than all other models

1. **Background and motivation**

According to Chao et al. (2002) demonstrated that the logistic regression model is a powerful model to use when the outcome variable is dichotomous in nature. The main limitations of this model is it considers only the additive nature of the variables. In health science there are situations which involves more than one dependent variables and each dependent variable simultaneously depends on remaining dependent variables in addition to the repressors in the form of system of equations. Even though machine learning models are effective and efficient classifier for future decision making, most of Machine learning models explains only the direct effect of single study variable on the independent variables and cannot be used to explain the intermediate effects. To identify and model the direct and indirect effect between variables simultaneously and effectively is a biggest challenge for researcher. Tanya et al (2010) advised that structured equation modeling can provide a new perspective on analyzing the data in medical and health science.

In this paper, an empirical study is carried out to show the effectiveness of path modeling to handle complex nature among variables to predict survival of ICU patients.

Objectives

* Identifying the risk factors associated with ICU mortality
* Predicting the probability of survival of ICU patients
* Selecting the best classifier for predicting the vital status
* A predictive model for studying the simultaneous effect of these risk factors for ICU mortality.

1. **Methodology**

**3.1 Logistic Regression model**

Logistic regression analyses the relationship between one or more independent variables and a categorical dependent variable and estimates the probability of occurrence of an event by fitting data to a logistic curve. Logistic regression is classified into two types namely binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. The logistic regression model is given by

WhereX1= x1 ,…,Xk = xk)

**3.2 Decision Tree**

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leafnode (or terminalnode) holds a class label. The topmost node in a tree is the root node. Given a tuple ***X***, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction/predicted values for that tuple.

**3.3 Random forest classifier**

Random forest is a supervised learning algorithm which is used for both classification as well as regression because of its simplicity and high accuracy. Random forest algorithm consists of multiple decision trees on study samples and then gets the prediction from each of them and finally selects the best prediction by means of voting. It is an ensemble method which is better than a single decision tree. It reduces the over-fitting by averaging the result.

**3.4 XGBoost classifier**

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for regression and classification predictive modelling problems. Ensembles are constructed using decision tree models. Trees are added one at a time to the ensemble which helps to correct the prediction errors made by prior models. XGBoost have the base learners when all the predictions are combined, bad predictions cancels out and better one sum up to form final good predictions. The XGBoost is having a tree learning algorithm as well as linear model learning, and because of that, it is able to do parallel computation on a single machine.

**3.5** **K-Nearest Neighbor classifier**

K-nearest neighbors is based on learning by analogy, that is by comparing given test tuple with training tuple that similar to it. When given an unknown tuple, a K- nearest neighbor algorithm search the pattern space for the k-training tuple that are closest to the unknown tuple. Closeness is measured in terms of distance matric. K-nearest neighbor method can also be used for prediction, that is, to return a real valued prediction for a given unknown tuple. In this case, algorithm returns the average value of the real valued labels associated with the K-nearest neighbors of the unknown tuple.

**3.6 Support Vector Machine**

The objective of SVM is to look for the optimal separating hyper plane between classes. The points lying on classes’ boundaries are called support vectors, and the in-between space is called the hyper plane; when a linear separator is not able to find a solution, data points are projected to a higher-dimensional space, where the previous nonlinearly separable points become linearly separable, using kernel functions. The whole task can be formulated as a quadratic optimization problem that can be solved with exact techniques. SVM aims at maximizing the margin between the support vectors and the hyper plane.

**3.7 Linear Discriminant Analysis**

Linear discriminant analysis aims on the association between multiple independent variables and categorical depend variable by forming composite of independent variables. This method discriminates between two or more preexisting groups and also derives the classification model for predicting the group membership of new observations. Fisher Linear Discriminant analysis assumes independent variables normally distributed with equal covariance matrix

Where

According to Fisher Linear Discriminant rule

Where

**3.8 Path Modeling**

Path model is a special case of structure equation model. Structure equation model is a combination of multiple regression model and factor analysis whereas Path model deals only with measured variables. Path Modeling is used to examining the effect of the many independent variables on the one or more dependent variables. It is similar to regression analysis, but it explains both direct and indirect effects between the variables simultaneously and effectively. The sum of all direct and indirect relationship gives total association of a variable.

Rex B. Kline (2005) and Daire Hooper, et al (2008) suggested following popular fit statistics and recommended respective cut-offs that stipulate a good fit,

|  |  |
| --- | --- |
| Model Chi-square statistic | p-value > 0.05 |
| (Adjusted) Goodness of fit | GFI ≥ 0.95 and AGFI ≥ 0.90 |
| (Non) Normed-Fit Index | NFI ≥ 0.95 AMD NNFI ≥0.95 |
| Comparative Fit Index | CFI ≥ 0.90 |
| RMSE of Approximation | RMSEA < 0.08 |
| Root Mean Square Residual (Standardized) | SRMR < 0.08 |

1. **Empirical Study**

Our study contains 102 patients who are admitted to ICU (Intensive Care Unit ) along with 16 variables. Here the variables include Age, Heart Rate, pH from initial blood gases, Carbon Dioxide, Hydrogen Bicarbonate, Creatinine, Level of consciousness at ICU admission, Sex, Cancer part of present problem, Kidney injury, Stroke, Heart failure, Hepatic failure, Respiratory, Brain injury and Vital Status. R software and LISREL 8.80 version are used to analyse the data.

1. **Results and Discussions**

Study dataset contains 16 variable which will influence the survival of ICU patients. Since our dataset contains dichotomous variables, we fit logistic regression model to identify significant variables responsible for the mortality of the ICU patients.

Table 1. Estimated regression coefficients, Standard errors, Wald test statistic with the corresponding P-value

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) | Conclusion |
| (Intercept) | 1.64 | 0.78 | 2.11 | 0.04 | Significant |
| age | -0.01 | 0.00 | -3.98 | 0.00 | Significant |
| HR | 0.00 | 0.00 | -1.93 | 0.06 | Significant |
| pH | -0.03 | 0.02 | -1.99 | 0.05 | Significant |
| CO2 | -0.01 | 0.02 | -0.30 | 0.77 | Insignificant |
| HCO3 | 0.01 | 0.04 | 0.29 | 0.78 | Insignificant |
| creatinine | 0.00 | 0.03 | 0.07 | 0.94 | Insignificant |
| LOC | 0.06 | 0.01 | 4.91 | 0.00 | Significant |
| Sex | 0.03 | 0.09 | 0.33 | 0.74 | Insignificant |
| cancer | -0.66 | 0.25 | -2.63 | 0.01 | Significant |
| kidney injury | -0.17 | 0.18 | -0.93 | 0.36 | Insignificant |
| stroke | -0.34 | 0.18 | -1.90 | 0.06 | Significant |
| heart failure | -0.31 | 0.19 | -1.70 | 0.10 | Significant |
| hepatic failure | -0.35 | 0.20 | 0.20 | 0.08 | Significant |
| respiratory | -0.30 | 0.16 | 0.16 | 0.08 | Significant |
| brain injury | -0.11 | 0.20 | 0.20 | 0.60 | Insignificant |

From the above table we can observe that the variables age, Heart Rate, pH from initial blood gases, Level of consciousness at the time of ICU admission, Cancer part of present problem, Stroke, Heart failure, Hepatic failure and Respiratory are significant for the survival of the ICU patients.

Now to test the overall significance of the model, we will remove all the insignificant variables from the dataset and the model is fitted with only the significant variables. Then the overall significance of the model is obtained through likelihood ratio test.

Table 2. To test the overall significance of the model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Df | LogLik | Df | Chisq | Pr(>Chisq) |
| 2 | -71.601 |  |  |  |
| 11 | -22.362 | 9 | 98.479 | < 2.2e-16 \*\*\* |

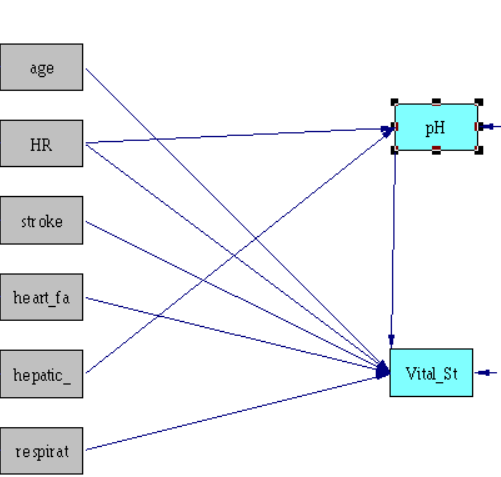
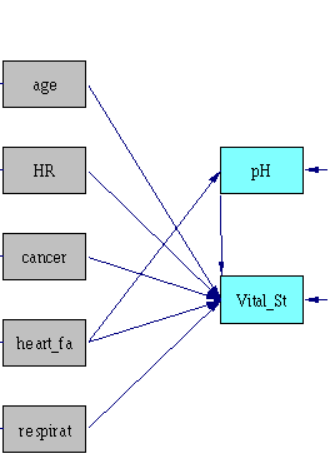
Observe that in table 2, P-value of logistic model fit statistic is less than 0.05. So we reject our null hypothesis and conclude that the fitted model is a good fit.

In most of the cases we ignore the complex nature that present between the variables by just considering the direct relation. But in health science related study, it is necessary to analyse the complex nature between the variables. We have to consider both direct and indirect effects on dependent variable by other variables. For this purpose if we go with regression analysis, we have to carry out multiple regressions to include indirect effects in our study. In this paper, path modeling with a mediating effect is established to study the effect of mediating variable on the survival of ICU patients. Here pH value is considered as mediating variable and we have fitted 3 models. Their statistics and path diagram is given below.

Table 3. Goodness of fit Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | P-value | RMSEA | SRMR | NFI | CFI | GFI |
| **1** | **0.97663** | **0.005** | **0.013** | **0.99** | **0.99** | **0.99** |
| 2 | 0.31081 | 0.045 | 0.04 | 0.96 | 0.98 | 0.98 |
| 3 | 0.73485 | 0.007 | 0.027 | 0.95 | 0.97 | 0.97 |

**Path diagrams**

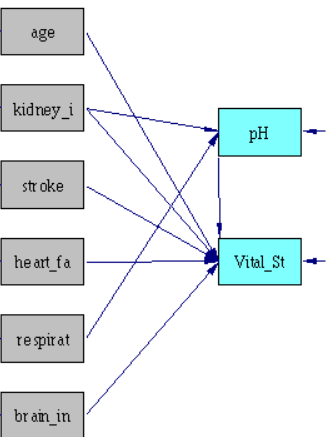


Table 4. Comparison of performance Path modeling with other models based on Accuracy and Recall.

|  |  |  |  |
| --- | --- | --- | --- |
| **Methods** | **Accuracy** | **Sensitivity** | **Specificity** |
| Logistic Regression | 78.92 | 91.48 | 78.57 |
| Random Forest | 79.92 | 92.35 | 78.87 |
| Linear Discriminant Analysis | 75.47 | 90.28 | 76.42 |
| Decision Tree | 76.85 | 63.64 | 85.71 |
| Xgboost | 86.65 | 90.9 | 83.57 |
| K-Nearest Neighbour | 84.72 | 81.82 | 85.59 |
| Support Vector Machine | 78.42 | 70.82 | 84.65 |
| Path Modeling | 90.45 | 90.91 | 87.41 |

Above table shows that performance of Path modeling is better as compared to all other fitted models based on accuracy, sensitivity and specificity measures.

**Conclusion**

In health science, machine learing methods extensively used to identify the hidden pattern through predictive classifier which helps in future decision making. Identification of significant factors and correct detection of the chance of survival of ICU patients using an efficient classification model will reduce the burden of doctors as well as avoid great loss. The results of the logistic regression model shows that Age, Heart Rate, pH from initial blood gases, Level of consciousness at the time of ICU admission, Cancer part of present problem, Stroke, Heart failure, Hepatic failure and Respiratory are the significant factors for the survival of ICU patients. Most of the classification models considers only the additive nature and fails to explain the both direct and indirect effect among the variables. This study shows the effectiveness of path modelling to handle complex nature among variables to predict survival of ICU patients. The result shows that path modelling outperforms all other classification models under considerations. The results from path modeling can be used to predict survival of ICU patient accurately which helps patient family to decide whether to go for further treatment or not.

**REFERENCE**

# [Tanya N Beran](https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-3-267#auth-Tanya_N-Beran), [Claudio Violato](https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-3-267#auth-Claudio-Violato) (2010) “Structural equation modeling in medical research: a primer “ Beranand Violato BMC Research Notes 2010, 3:267 http://www.biomedcentral.com/1756-0500/3/267

# Qais Abdulqader (2017) “Applying the Binary Logistic Regression Analysis on The Medical Data”[Science Journal of University of Zakho](https://www.researchgate.net/journal/Science-Journal-of-University-of-Zakho-2410-7549) 5(4):330

DOI:10.25271/2017.5.4.388

1. Senthilkumar and Sachithanantham “A Study On The Incidence Of Tuberculosis Using Binary Logistic Regression” (2018)

International Journal of Recent Scientific Research Vol. 9, Issue, 4(I), pp. 26159-26169, April, 2018

1. [Wendong Ge](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ge%20W%5BAuthor%5D&cauthor=true&cauthor_uid=30815086), [Jin-Won Huh](https://www.ncbi.nlm.nih.gov/pubmed/?term=Huh%20JW%5BAuthor%5D&cauthor=true&cauthor_uid=30815086), [Yu Rang Park](https://www.ncbi.nlm.nih.gov/pubmed/?term=Park%20YR%5BAuthor%5D&cauthor=true&cauthor_uid=30815086), [Jae-Ho Lee](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lee%20JH%5BAuthor%5D&cauthor=true&cauthor_uid=30815086), [Young-Hak Kim](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kim%20YH%5BAuthor%5D&cauthor=true&cauthor_uid=30815086) (2018) “An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units”

[AMIA Annu Symp Proc.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371274/) 2018; 2018: 460–469

1. [Sonu Subudhi](https://www.nature.com/articles/s41746-021-00456-x#auth-Sonu-Subudhi), [Ashish Verma](https://www.nature.com/articles/s41746-021-00456-x#auth-Ashish-Verma),,[Ankit B. Patel](https://www.nature.com/articles/s41746-021-00456-x#auth-Ankit_B_-Patel), [C. Corey Hardin](https://www.nature.com/articles/s41746-021-00456-x#auth-C__Corey-Hardin), [Melin J. Khandekar](https://www.nature.com/articles/s41746-021-00456-x#auth-Melin_J_-Khandekar), [Hang Lee](https://www.nature.com/articles/s41746-021-00456-x#auth-Hang-Lee), [Dustin McEvoy](https://www.nature.com/articles/s41746-021-00456-x#auth-Dustin-McEvoy), [Triantafyllos Stylianopoulos](https://www.nature.com/articles/s41746-021-00456-x#auth-Triantafyllos-Stylianopoulos), [Lance L. Munn](https://www.nature.com/articles/s41746-021-00456-x#auth-Lance_L_-Munn), [Sayon Dutta](https://www.nature.com/articles/s41746-021-00456-x#auth-Sayon-Dutta) (2021) “Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19” NPJ Digit Med. 2021 May 21;4(1):87. doi: 10.1038/s41746-021-00456-x.
2. Suojeng Huer,Ryoung – Eun Ko,JunSang Yoo (2021) “A Machine Learning-Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study” . [JMIR Med Inform.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8367129/) 2021 Jul; 9(7): e23401.

Published online 2021 Jul 26. doi: [10.2196/23401](https://dx.doi.org/10.2196%2F23401)

1. Siyi Yuan,Yunbo Sun,Yun Long,Huawi He (2021) “Using Machine Learning Algorithms to Predict Candidaemia in ICU Patients With New-Onset Systemic Inflammatory Response Syndrome” [Front Med (Lausanne).](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8416760/) 2021; 8: 720926.

Published online 2021 Aug 19. doi: [10.3389/fmed.2021.720926](https://dx.doi.org/10.3389%2Ffmed.2021.720926)

1. Suojeng Huer,Ji Young Min, Won Chul Cha, Kyunga Kim (2021) “Development and Validation of Unplanned Extubation Prediction Models Using Intensive Care Unit Data: Retrospective, Comparative, Machine Learning Study”

[J Med Internet Res.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8387891/) 2021 Aug; 23(8): e23508.

Published online 2021 Aug 11. doi: [10.2196/23508](https://dx.doi.org/10.2196%2F23508)

1. Jialin Liu,Siru Liu,Kunchung Hu (2021) “Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model”

[*PLOS ONE*](https://econpapers.repec.org/article/plopone00/), 2021, vol. 16, issue 2, 1-11