**MORTALITY PREDICTION:**

**To Be, Or Not To Be**

[[1]](#footnote-1)

Akash Verma Shreya Goyal Shridhar Kumar Thakur

Department of Computer Science Department of Computer Science Department of Computer Science

Bhilai Institute Of Technology National Institute of Technology Galgotias University

Durg Jalandhar Greater Noida

akashverma.av98@gmail.com shreyag.2199@gmail.com thakur.sridhar1@gmail.com

Archit Gupta Indrajeet Gupta

CSE Department Department of Computer Science

ABES Engineering College,Ghaziabad Bennett University

architg380@gmail.com indrajeet.gupta@bennett.edu.in

***Abstract***— **In these past years, there is a increase in interest towards data mining, artificial intelligence and machine learning concepts to improve hospital performance. In some hospitals, they want to improve their statists by decreasing the number of patients dying in the hospital. The Research is focused in mortality, prediction of measurable outcomes, including risk of complications & length of hospital stay. The duration spent in the hospital plays an important role both for patients & healthcare providers , influenced by numerous factors. In particular LOS(length of stay) in critical care is of great importance, both to patient experience as well as the cost of care, and is influenced by the complex environment factors of the Hospitals. LOS is a parameter which is used to identify extremity of illness & health related resource utilisation. This paper examines the case of either a patient survives or dies in the range of length of stay in the hospital. It also anchors the analytical methods for length of stay and mortality prediction.**

*Index Terms*—Machine Learning Algorithm,Featue Scaling, Feature Extraction, Neural Networks, Logistics Regression, Linear SVM.

# **INTRODUCTION.**

Life is the most valuable thing in the world. Whatever is in the world is due to life. Due to some consequences or situations, a person may got injured or get ill with diseases.These factors are one of the major causes of Death. In hospital doctors try to treat and cure these diseases and problems. With the advancement of technology the chances of survival of a patient has been increased. One of the biggest emerging technology is the Artificial Integillence. Machine Learning Model can help boost up predicting critical conditions and helps to take prevention and save a life.

**Mortality Prediction:To Be, Or Not To Be** is a competition Hosted by ***Codalab***for the learning and enrichement of students in the machine learning field but the problem is very closely related to real life.

Everyday, doctors and nurses collect a lot of information about patients by asking questions and using adapted tools utilizing stethoscope, syringe, handheld sensor, paper reports, etc. The dataset contains many features such as the heart rate, respiration rate, glucose level,either there is presence of particular disease or any particular symptoms,etc. The training set has 79,999 rows and 342 features without an output column. The output column is available as a separate train\_solution file which specifies the status of the patient about whether it survives or not. Out of the 342 features in the data, we have 7 categorical variables, 333 numerical variables, and 2 date variable. Through this very project, we have encountered many challenges and risks that we tried to mitigate. Here are the risks and challenges we faced in our project:

1. **Missing values:**

The dataset from Codalab site contains missing values. For example, the Marital Status feature of the categorical feature contained 2 missing values which cause difficulty in further processing. We also had a lot of NAN values in our numerical data. We mitigated the missing categorical values risk by replacing them with the most occurred value of the feature. The numerical missing data was replaced by the mean of the column.

1. **Different Values:**

The very next problem in our dataset was the different value in different categories. The dataset contains many features that contain similar representation of data but there are 3 different categories as float, integer, and object in which we need to convert string values to numeric. We have tried to avoid this risk by identifying these features and dropping the unnecessary features. Also, there was a risk of overfitting of our data during training by including all the features in our analysis.

1. **Class Imbalance:**

The dataset provided to us has extreme class imbalance, with 77,203 patients surviving and only 2797 patients dying. So, training and testing a model with these data surely leads to overfitting. In this scenario, we want to increase the minority class in our dataset that would essentially result in effective training and building of model. At the same time, we want to have lesser majority class, i.e. we like to consider lesser number of patients surviving. This would reduce the unnecessary overfitting. To mitigate the issue of class imbalance, we have preferred undersampling and oversampling based approaches.

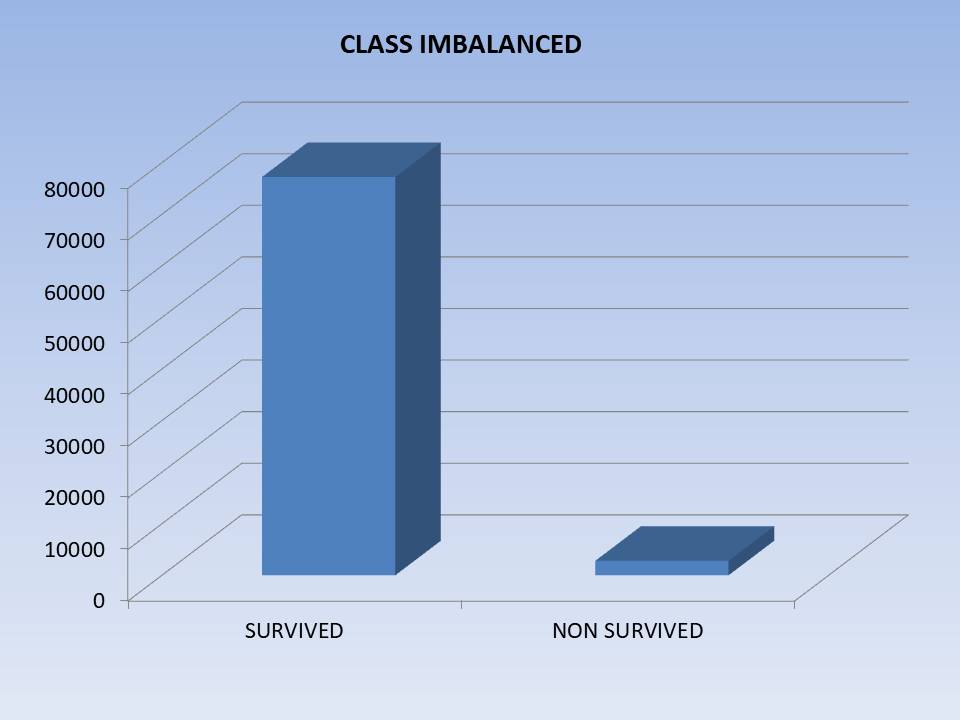


Figure1: Class Imbalance

Most common approach for this problem is classification using machine learning algorithm like Random Forests, Logistics Regression, etc. Random Forest is an ensemble learning method for  regression,classification & other tasks that operates by constructing multiple  decision trees during training time and output the class which is the mode of the classes (for classification) or mean prediction (in case of regression) of the individual trees. Random forest is useful for datsets that leads to overfiting. Another approach for solving the problem was classification using Deep Neural Network. Deep Neural Network is considered to be one of the most efficient classifier. In hidden layers using Relu Functions and using a softmax function at the outer layer, it showed accuracy around 605 percent .Other approaches for the same problem were Gaussian Naïve Bayes, Perceptron Model. The previous approaches required a lot of computational ability of the system and also the accuracy was not up to the mark.

Our approach for this very Project is a machine learning algorithm called Logistics Regression.

# **RELATED WORK**

**Mortality Prediction: To Be, Or Not To Be** is basically a Binary Classification problem. So there has been various approaches to tackle this problem. Some have chosen neural networks while others chose various Machine learning Classifier to have a better result. Neural Networks are considered to be one of the best Classifiers. As our dataset is nonlinear, we need to use deep neural network to introduce non linearity. This is one of the simplest approaches to our Problem but it has not shown much accuracy. The highest score we got using this approach was around 0.6049. And also tuning of a deep neural networks requires a good amount of time. Another approach using machine learning algorithm is Random Forest. Random forest is a ensemble learning algorithm which creates a multiple decision trees to predict the class. By tuning the random forest classifier we increased our score around 0.7166. Random forests have the main disadvantage as their complexity. They are much harder & time consuming to construct in comparison to decision trees. Furthermore its tuning is difficult and Overfitting can easily occur.

We have opted Logistics Regression Machine Learning Model, which is providing us better results than other neural and machine learning algorithms.

The implementation of the model supports the features of the scikit-learn and accuracy score implementations.

# **Methodology**

## Problem Statement

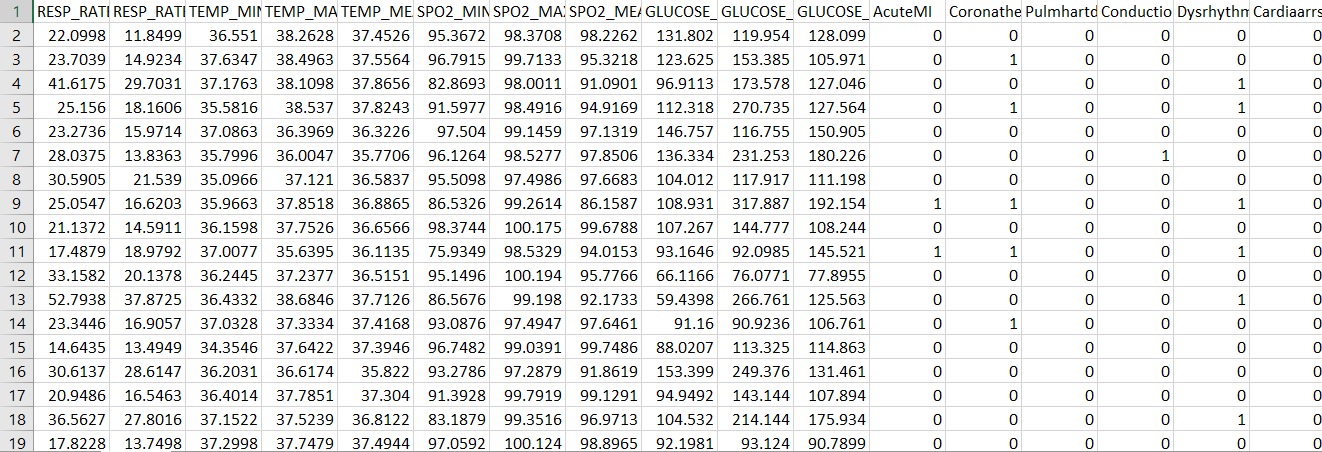
In this problem, We have to predict that whether a patient survives or not during his/her stay in the hospital. To be able to predict we use Linear SVM as well as Logistics Regression machine learning model using the training set as well as testing set.

## Dataset Description

The Dataset has been provided by the Codalab with reference and credits to Isabelle Guyon, Kristin Bennett, Andrew Yale, Adrien Pavao, Thomas Gerspacher. The Dataset contains 79,999 patient entries with 342 features and a multi label status with three different values. The Testing Dataset contains 20,001 values.

## Preliminary Data Analysis

The first task for the project was to explore the dataset and try to establish relationship between different features of dataset and the labels & try to exclude those who were not affecting our labels. There was also a major problem as features like marital status,insurance had many missing data points, so We filled those missing data points with the mean and median(as required) of the respective feature in that particular column. Also, the labels and training datasets were given separately so we required to link all the datasets to combine and process so we can actually visualize the data.

 Figure2: Dataset Description

## Data Visualization

We need to explore the data as it is carrying so many features so we tried to find the correlation between the features and remove redundancy in our dataset. This reduced our features size but helped in decreasing the dataset length.

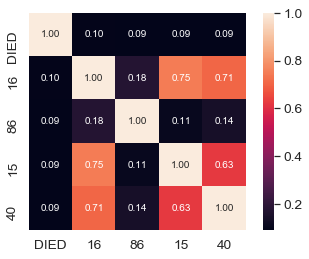


Figure3: Correlation Matrix

We replaced the missing values with mean of the column. We extracted the important features with unique categories which reduced our features from 342 to 162. We further required to put dummies for string variables which increased our features to 211.

We further did scaling of the data with min-max scalar. We use SMOTE(Synthetic Minority Over Sampling Technique) to handle the imbalanced data.

## Model selection and training

For further processing we dropped the unnecessary features to ease our training procedure. We reduced our feature size to 212.Now we have to choose the model to train our dataset. We tried with different models to fulfil our need. We tried with 5 different models and tried to tune them to maximize our result.

Table1: Description of Models

|  |  |  |
| --- | --- | --- |
| **Model** | **Description** | **Score** |
| Neural Network | 4 dense hidden layer with input size 211 and adam optimizer | 0.6049 |
| Gaussian Naïve Bayers | P(h)d = (P(d|h)\* P(h)) / P(d) | 0.6124 |
| Perceptron | Learn the weights to get the function | 0.7300 |
| Linear SVM | Classification by finding the hyper plane | 0.7513 |
| Logistics Regression | Estimating the parameters using a logical function | 0.7576 |

We first tried with Gaussian Naïve Bayes. It gave score around 0.6124. For Perceptron Model, we got score around 0.73 on drivendata.

For Deep Neural Network, we got the score of 0.6049.For Linear SVM model we increased our score to 0.7513.We got the best result from the Logistics Regression which gave a score of 0.7564.

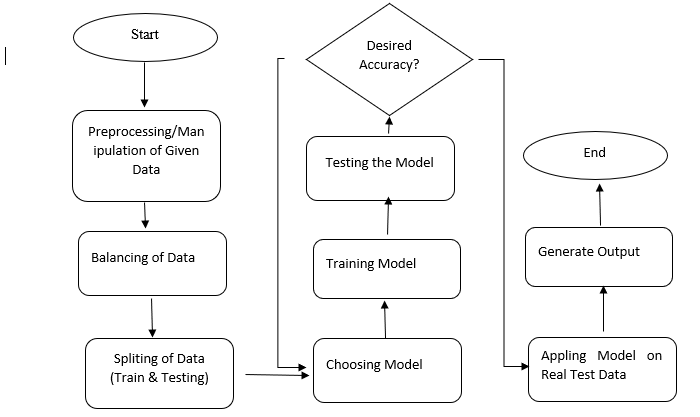


Figure4: Flowchart of our methodology

# **Experimental Results**

We had tried in 5 different algorithms models in that Gaussian Naive is giving 0.6124 accuracy and then the Perceptron showing the accuracy 0.73 and then the Linear SVM showing the accuracy 0.7513. Finally we had tried in the Logistics Regression, its shows the best accuracy that is 0.7564 so we take the Logistics Regression model has the best trained model for the “**Mortality Prediction: To Be, Or Not To Be**”. Below is the result we got from our best solution.

Table2: Result of Our best solution

|  |  |
| --- | --- |
| Model | Value counts(stats of dies) |
| Logistics Regression | 5269 |
| Linear SVM | 3062 |
| Linear SVC | 3058 |

Here is the graphical representation of accuracies which we got through various models:

Figure5: Accuracies of different models

# **Conclusion**

The Dataset provided by Codalab Website for prediction of the mortality of patient in terms of Numerical and Categorical values needs some repairs. We preprocessed the data and trained our dataset with five models (Support Vector Machine, Decision Trees, Deep Neural Network, Gaussian Naïve Bayes, Logistics Regression). Among all the models Logistics Regression gave us the best score of 0.7564.In our opinion this model will be helpful in predicting the condition of patient. We can improve the score by implying ensemble of 2-3 models and get a better result.

**References**

.

1. 1.Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults. Chest. 1991;100(6):1619–1636.
2. 2.  Che Z, Purushotham S, Cho K, Sontag D, Liu Y. arXivpreprint arXiv:1606.01865. 2016. Recurrent neural networks for multivariate time series with missing values
3. 3.  Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today’s critically ill patients. Crit Care Med. 2006 May;34(5):1297–1310.
4. 4.  Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med. 1985 Jul;13(7):519–525.
5. 5. Sharma A, Shukla A, Tiwari R, Mishra A. Mortality Prediction of ICU patients using Machine Leaning: A survey. Paper presented at: Proceedings of the International Conference on Compute and Data Analysis.2017.
6. [https://codalab.lri.fr/competitions/566#participate-get\_data](https://codalab.lri.fr/competitions/566)
7. Sharma A, Shukla A, Tiwari R, Mishra A. Mortality Prediction of ICU patients using Machine Leaning: A survey. Paper presented at: Proceedings of the International Conference on Compute and Data Analysis.2017.
8. <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>

1. [↑](#footnote-ref-1)