Prediction of Mortality Rates Post Thoracic Surgery

Tapas Bhadra[1] Shachi Sonar[2] Samruddhi Zagade[3] Prof. Sanjay Vidhani[4]
[1,2,3,4]Department of Information Technology
[1,2,3,4]K. J. Somaiya College of Engineering, India

Abstract— Cancer has been one of the biggest enemy to human health, with its insidious nature, and lack of available treatment. Termed to be the leading cause of death, responsible for an estimated 9.6 million deaths in a year. Amongst various types of cancer, we have narrowed our focus on Thoracic category of this disease. Tons of Lung Cancer patients every year undergo thoracic surgery but due to high mortality rate and less life expectancy the survival after one year is uncertain.

Keywords: Thoracic Surgery, Lung Cancer, Logistic Re-Gression, Random Forest Classifier

I. INTRODUCTION

CANCER has been one of the biggest enemy to human health, with its insidious nature, and lack of available treatment[7]. Termed to be the leading cause of death, responsible for an estimated 9.6 million deaths in a year. Amongst various types of cancer, we have narrowed our focus on Thoracic category of this disease, one of the most popular surgeries done on the lungs. The thorax which stretches from the oesophagus to the diaphragm contains the different organs and parts in between including the lungs. Many patients suffering from cancer of the thoracic organs such as lung cancer often have to undergo the thoracic surgery. But, many a times, the surgery can lead to further complications which may lead to death. Hence, it is necessary for the doctor to study the patient and make a decision on whether the patient must undergo the surgery or not. In certain situations, the patient might have a longer lifespan without having to undergo the surgery. There are various attributes or features that lead to generation of cancer cells in a human body making it extremely difficult to attribute the issue to a single feature. We can take an example of a person, who has never smoked a cigarette in his/her life, dying a month after surgery, and a chain smoker who survived the surgery and is living a long and healthy life. In this situation, we have to broaden our scope to a wide variety of reasons that can lead to mortality after the surgery. What we aim at achieving is to build a model that can aid the doctors in their decision making process of predicting whether the patient must undergo the surgery or not.

Our project focuses on the following:-

1) Feature engineering on a database consisting of 16 features and identifying the right attributes that contribute the most in the decision making.
2) Building and comparing the accuracy, and precision of two prediction models.
3) Developing a system that can be used by the doctors to make decisions regarding the whether the patient should go ahead with the surgery or not.

II. LITERATURE SURVEY

Work in this field has been undertaken by doctors from various fields in order to find any metric that would help them better diagnose patients. In one study[4], a Preoperative Score was calculated to predict the postoperative mortality rate. The aim was to diagnose a surgical risk factor purely based on previous information. The doctors observed 24 factors include age and the type of surgery to be conducted. The study used a multivariate logistic model, and gave a fairly accurate result.

Reducing the type of surgery only concerned with the lungs, Abeer S.[5] performed a compartive study of 4 machine learning algorithms, namely, Simple Logistic Regression, Support Vector Machine, Multivariate Logisitic Regression and Multilayer Perceptron learning. They used boosted SVM to get better results.

The background study shows that prior to the data mining and statics era the General Thoracic Surgery data(GTSD)[9] had risk models that had predictors which could predict the survival of a patient undergoing thoracic surgery only up to 30 days. With the advancement in technology and data mining it is now possible to find the long-term predictors of life expectancy of such patients.

In another paper[8], Naive Bayes along with j48 was used on their data set to improve accuracy.

III. SYSTEM DESIGN

A. Data set

The Thoracic Surgery Data[3] data set used for our project was elicited from UCI Machine Learning Repository[2]. The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung re-sections for primary lung cancer in the years 2007-2011[3]. There are 470 instances of multivariate data with over 16 attributes and 1 outcome column. This outcome divides the set of training data into two classes: class1 - ”true” denoting death within one year after surgery and class2 -”false” denoting the survival of patient one year post the surgery. The attributes show the status of the patients health before undergoing the surgery. The data set was used for training and testing purpose of our model.

Data-Set Description
• DGN: Diagnosis - specific combination of ICD-10 codes for primary & secondary as well multiple tumours(DGN3,DGN2,DGN4,DGN6 ,DGN5,DGN8,DGN1)
• PRE4: Forced vital capacity - FVC (numeric)
• PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
• PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
• PRE7: Pain before surgery (T,F)
• PRE8: Haemoptysis before surgery (T,F)
• PRE9: Dyspnoea before surgery (T,F)
• PRE10: Cough before surgery (T,F)
• PRE11: Weakness before surgery (T,F)
• PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
• PRE17: Type 2 DM - diabetes mellitus (T,F)
• PRE19: MI up to 6 months (T,F)
• PRE25: PAD - peripheral arterial diseases (T,F)
• PRE30: Smoking (T,F)
• PRE32: Asthma (T,F)
• AGE: Age at surgery (numeric)
• Risk1Y: 1 year survival period - (T)rue value if died (T,F)

Fig. 1: Data Set Attribute Description

B. *Pre-processing and outlier analysis*

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing[10].

There are 470 instances in the data set with 70 ”true” class patients and 400 ”false” class patients. There are no missing values that need to be handled. The column having Boolean ’F’ and ’T’ string value were converted to 0 and 1 for ease of handling. The column with ID was redundant and hence was removed. The columns with redundant sting and number values were converted to only number values with int datatype for better handling. The names of the columns were changed according to the semantics to human understandable format for ease during working with the data set

Outliers are considered as those points that are distant from their own neighbors in the data set[11]. The numeric column were considered during this stage which were PRE4, PRE5 and age. Graphical method of outlier detection using Box plots and scatter plots was implemented to find out 15 outliers in PRE5 and 1 in age. There were total 16 outliers and the entries containing the outliers were removed to get 454 instances for further classification process.

C. *Classifier*

Various Machine learning algorithms can be used in order to classify the patients. We will be considering 2 algorithms, namely, Logistic Regression and Random Forest Regression.

1) *Logistic Regression:*

Logistic Regression is a classifier which uses the logistic or Sigmoid function to assign a probability to the tuples in a dataset. Using Linear Regression would not be possible in our project due to its lack of flexibility. Linear Regression requires that a threshold be set in order for it to classify. Consider a case where the actual class of a tuple is malignant, the predicted value is 0.5 and the threshold set is 0.6. In this case the model will predict it as benign. from this it can be inferred that Linear Regression is not suitable for classification and hence Logistic Regression. The value ranges from 0 to 1

The hypothesis is given as

$$h(x) = \text{sigmoid}(Z)$$

The Sigmoid function is given as

$$y = 1/(1 + e(-x))$$

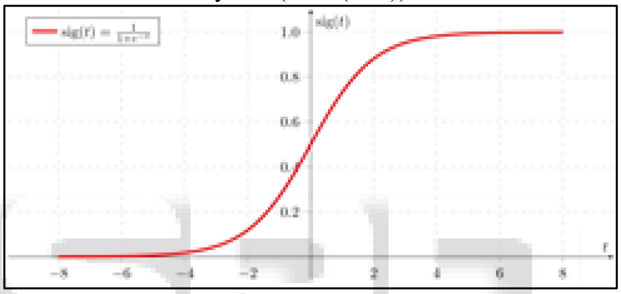

Fig. 2: Sigmoid activation Function

2) *Random Forest Classifier:*

The Random Forest Classifier is also known as the Random Decision Forest. It contains a large number of Decision Trees that work as an ensemble. Each Decision Tree classifies the sample of data and the class with the most number of votes becomes the prediction of the model. A collection of trees will outperform a single tree and hence gives a better result. It also helps from overfitting, as one tree can give a biased answer but a group of them will avoid overfitting.

IV. FEATURE ENGINEERING

With the abundant amount of features, there is a need to extract those features that have a higher impact on the decision making process. This will further lead to an increase in the accuracy of our model. Amongst the 15 features excluding Death 1year which is the dependent class that we want to predict, we have to extract those features that weigh more in the prediction process.

Out of the 454 patients, 69 patients did not survive 1 year after which accounts to 15.20% of the total data. In order to compare the two different patient classes, we segregated the data into two sets, live and dead, and compared their mean values. We can see the mean differences of all the attributes for live as well as dead patients in Fig2.

	Death 1yr Mean	Live 1yr Mean
Attribute		
FVC	3.195072	3.304597
FEV1	2.383188	2.540805
Performance	0.913043	0.774026
Pain	0.101449	0.051948
Haemoptysis	0.202899	0.124675
Dyspnoea	0.115942	0.044156
Cough	0.797101	0.677922
Weakness	0.246377	0.158442
Tumor_Size	2.014493	1.683117
Diabetes_Mellitus	0.144928	0.062338
MI_6mo	0.000000	0.005195
PAD	0.028986	0.015584
Smoking	0.898551	0.815584
Asthma	0.000000	0.005195
Age	63.333333	62.677922

Fig. 3: Mean Values of Live and Dead Patients

After normalizing the values, we plotted the %Mean Difference between the attributes that can be seen in figure3.

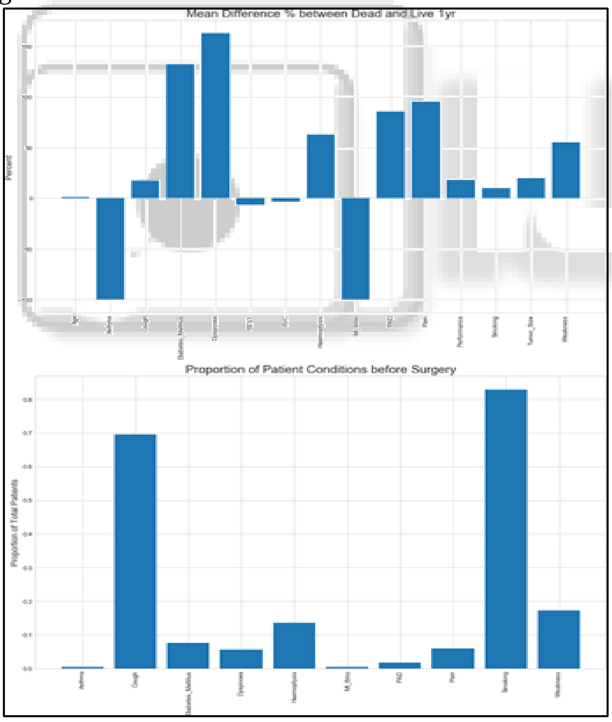

Fig. 4: % Mean Differene between Live and Dead Patients

Through the bar plot in Fig3, we can see that the most significant attributes for the patients who died are Dyspnoea, Diabetes Mellitus, Pain, PAD, and Haemoptysis in the de- creasing order namely. This indicates that these attributes were strongly present for those patients who died within a year after surgery. While, MI of 6 moths and Asthma showed negative results. After studying the proportion of patient conditions before surgery shows that almost 0 patients had Asthma and Mi of 6 months.

In order to analyze the other significant features that we obtained from the Mean difference plot, a Count Graph was used which shows the difference in the counts of the attributes, Fig5.

For Diagnosis, a last majority of patients are in category 3. The proportion of live to dead is the same throughout except for category 5. Here the number of dead patients is more than alive.

For Tumor Size, the categories 1 and 2 are the majority. The proportion of dead to live gently increases with the increase in tumor size.

For performance, categories 1,0 and 2 are in the order.Category

0 represents low death count which supports the Zubrod Scale where 0 is good and 2 is poor.

A. Hypothesis Testing:

In order to validate the correlations and observations that we have drawn, we used hypotheses testing to ascertain that the trends in the data were not just by chance.

1) Null Hypothesis:

The 1 year live and death patients have the same distribution and mean.

2) Test Statistic:

Mean difference between live and dead patients.

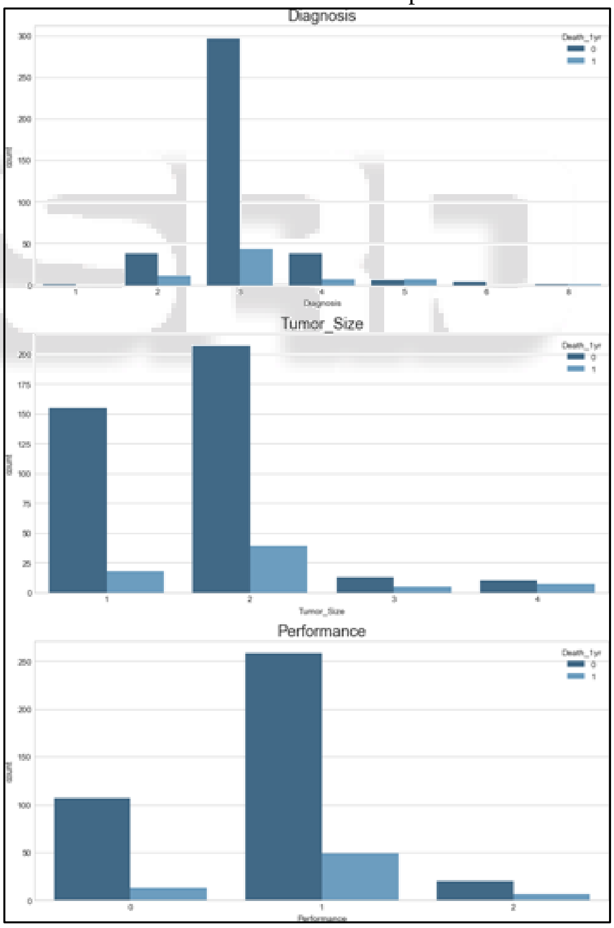

Fig. 5: Count Graph

3) Level Of Significance: 0.05

The attributes of interest are those that reject the null hypothesis. Out of all the attributes, 5 features satisfy this condition and reject the Null hypothesis. These features will have different distributions for live and dead patients and hence we are interested in them.

The features that rejected this hypothesis, with their mean difference for death within 1 year are given below

- Performace: 17.96%
- Dysnoea: 162.57%
- Cough: 17.58%
- Tumor Size: 19.69%
- Diabetes Mellitus: 132.49%

B. Summary of Results:

Amongst the features, Dyspnoea, Diabetes Mellitus, Pain, PAD, and Haemoptysis presented strongly. Most of the patients who underwent the surgery, smoked and showed symptoms of coughing. Majority of the patients were attributed to Diagnosis code 3. The dead to live ratio is in accordance with the Diagnosis codes. The data presents a trend for the number of dead patients with an increase in the Tumor size. For performance, the trend observed is, an increase in dead patients as performance on Zubrod Scale increases.

Hypothesis Testing revealed that the attributes are Performance, Dyspnoea, Cough, Tumor Size, Diabetes Mellitus.

V. OUTPUT ANALYSIS

We have separated our data into two inputs, one is 'x' which contains all our input parameters and the other being 'x2', which we have narrowed down through feature extraction.

X2 contains the following parameters - 'Performance', 'Dyspnoea', 'Cough', 'Tumor Size', 'Diabetes Mellitus' from our dataset. A comparison was done on our logistic regression model by using both x as well as x2 for our input.

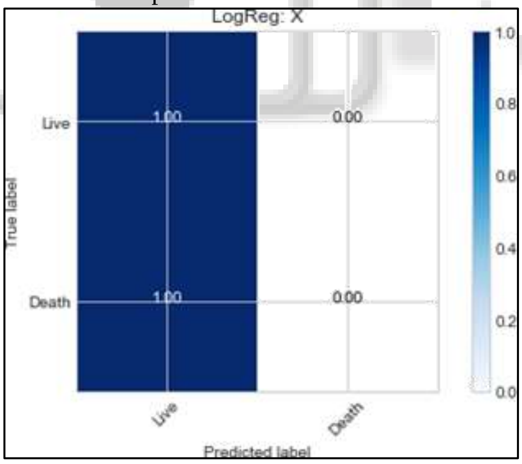

Fig. 6. Confusion Matrix of Logistic Regression on x with no class weight

On taking 'x' as our input, as expected in Fig. 6, there was poor performance. The model predicted that all of the patients would survive. The dataset being imbalaced due to the lower cases of death plays a role in this.

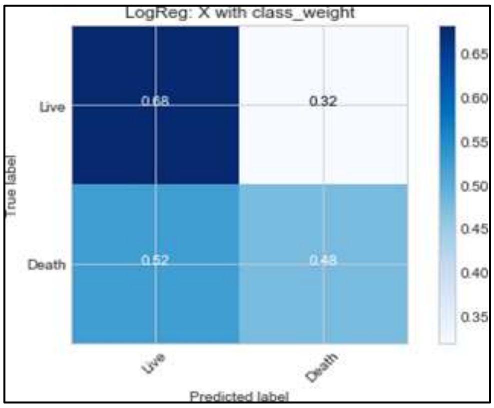

Fig. 7. Confusion Matrix of Logistic Regression on x with class weight

As we can see in the confusion matrix given by Fig. 7, with the addition of class weights, and a balanced data set, we are getting a much better precision of 0.18 with more realistic results. These improved results can be attributed to the class weight influenced along with the balanced data

When we took 'x2' as our input, without any class weight, as we can see in Fig. 8, there wasn't a drastic change in the performance of the model.

However, as seen in Fig. 9, by balancing our data and adding class weight, our precision has increased to 0.20.

In Fig. 12, we've taken input 'x2' and applied the random forest classifier to it. It's performance without class weights is almost similar to that of Logistic Regression.

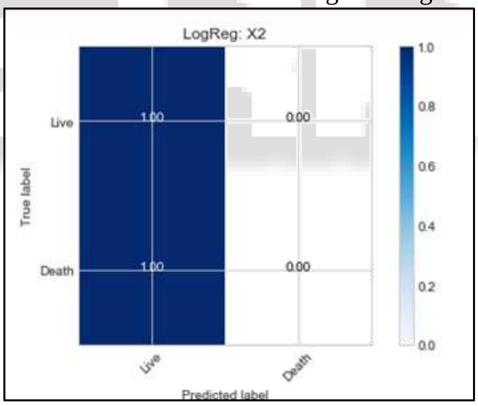

Fig. 8: Confusion Matrix of Logistic Regression on x2 with no class weight

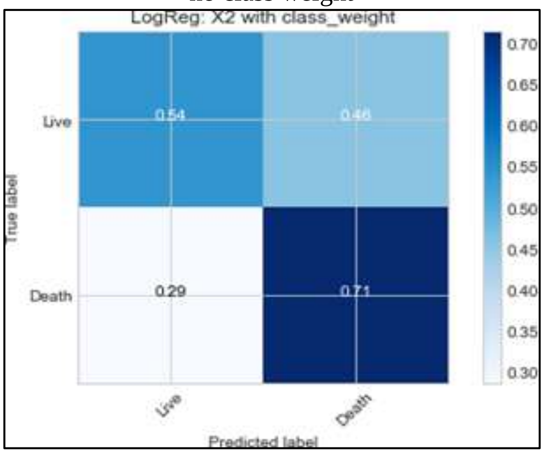

Fig. 9: Confusion Matrix of Logistic Regression on x2 with class weight

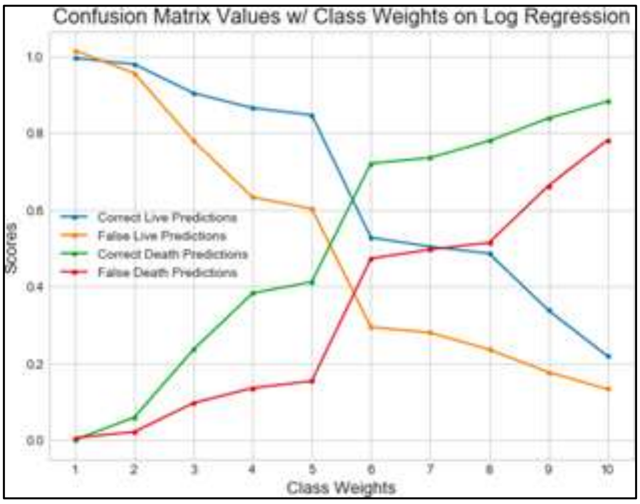

Fig. 10: Confusion Matrix with class weight on Logistic Regression

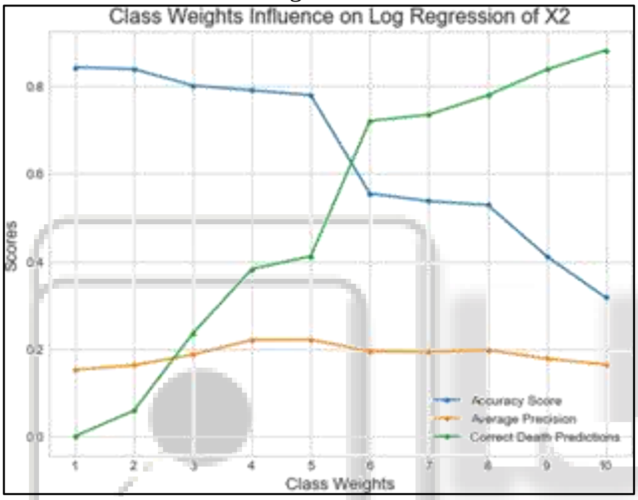

Fig. 11: Class weight influence on Logistic Regression

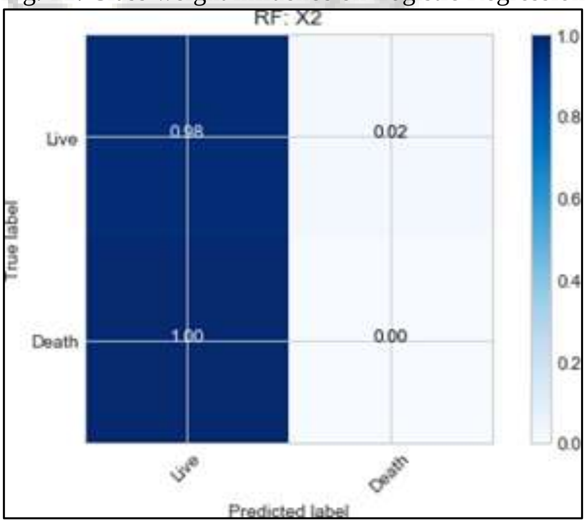

Fig. 12: Confusion Matrix of Random Forest Classifier on x2 with no class weight

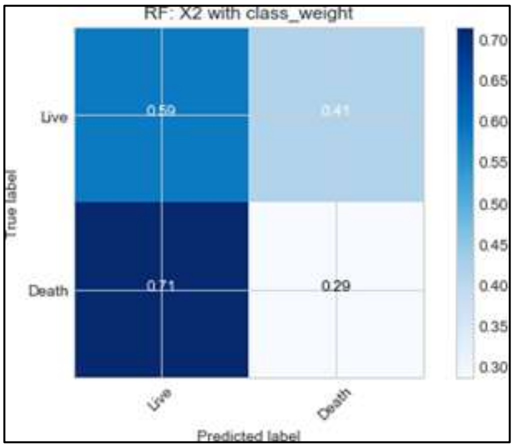

Fig. 13: Confusion Matrix of Random Forest Classifier on x2 with class weight

Even after balancing the dataset and adding class weights, we can see from Fig. 13, that Logistic Regression with class weights gave a better performance with precision 0.20 than Random Forest with precision 0.14.

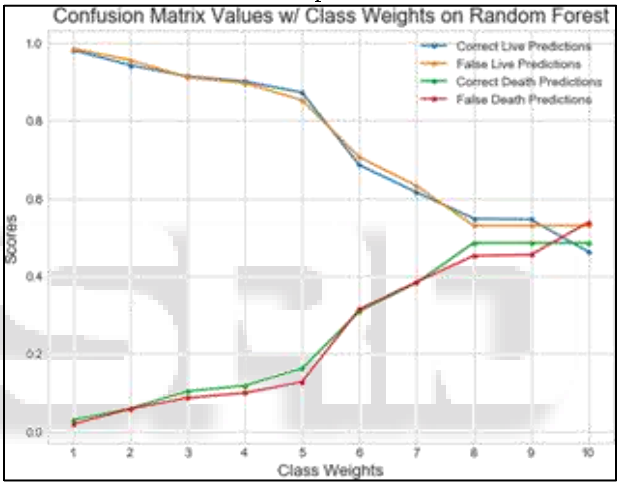

Fig. 14: Confusion Matrix with class weight on Random Forest

When we performed hyperparamater tuning for our Random Forest Classifier, it delivered a precision score higher than any other models (0.26). However, we do that see that the best parameters that were found through GridSearchCV, had no class weights. More combinations can be tried to give a better precision score.

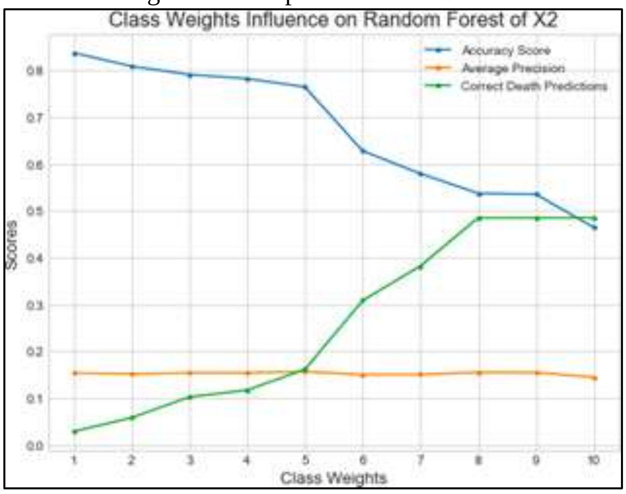

Fig. 15: Class weight influence on Random Forest

VI. Implementation

The implementation of this project encompasses:

- Creation of the Machine Learning model for predication of post-operative life expectancy of lung cancer patients who undergo thoracic surgery and a back-end controller to handle the data.
- Front-end Graphical User Interface to interact with the user.

A. ML Model and Backend

The model was trained using Random Forest classifier which gave the highest accuracy. Functions from Python library sklearn were used for the same. The .pki format model of was produced using pickle library. The Python Flask web framework was used in the back-end for handling HTTP requests and responses.

B. Graphical User Interface

A web-application was made as a Graphical User Interface to input the attribute values of the patient who's life expectancy is to be predicted. The web-app also has a secure registration- login system to ensure there is no unauthorised access and misuse of the application. Web technologies like HTML and CSS were used to design the web-app and PHP was used for data handling and validation. Relational Database system MySql database was used for user and patient database man- agement.

VII. Conclusion

Therefore, we have created a web application aimed for a professional use by doctors. Through this study, we came to know multiple factors that play a much greater role in pre- dicting mortality rate. Using these parameters, we compared Logistic Regression Model and Random Forest Classifiers with both the inclusion of class weights and balancing of data. Random Forest Classifier with hyperparamter tuning is most likely to give the best result.

After opting for the best model, we used that as the machine learning component of our web application. Using Flask for our backed, along with PHP and HTML, we were able to cre- ate our web application. This application with its intuitive User Interface and design aims to provide a comprehensive report of the mortality rate of the patient, given their preoperative underlying medical history. It's purpose is to give a statistical backing for decisions that a patient and their doctor can take knowing that they have all the background information and future predictions.

Acknowledgment

Firstly, we would like to thank our esteemed institution, K. J. Somaiya College of Engineering for supporting our research project. We would also like to extend our gratitude towards our mentor Prof. Sanjay Vidhani, Department of Information Tech- nology, KJSCE, for his guidance and motivation throughout the project. His constructive criticism, expertise and constant support has got the best out of us.

References

[1] H. Kopka and P. W. Daly, A Guide to LATEX, 3rd ed. Harlow, England: Addison-Wesley, 1999.
[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Reposi- tory [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[3] Marek Lubicz, Konrad Pawelczyk, Adam Rzechonek and Jerzy Kolodziej, Thoracic Surgery Data Data Set, 2013.[On-line].Available: http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
[4] Yannick Le Manach, M.D., Ph.D.; Gary Collins, Ph.D.; Reitze Rodseth, M.B.Ch.B., Ph.D.; Christine Le Bihan-Benjamin, M.D., M.Sc.; Bruce Biccard, M.B.Ch.B., Ph.D. Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and Validation, Anesthesiology 3 2016, Vol.124, 570-579
[5] Abeer S. Desuky, Lamiaa M. El Bakrawy Improved Prediction of Post- operative Life Expectancy after Thoracic Surgery, Published 2016
[6] Adam Abdulhamid, Ivaylo Bahtchevanov, Peng Jia. Life Expectancy Post Thoracic Surgery,Stanford University - CS 229.
[7] Urvay SE, Yucel B, Erdis E, et al. Prognostic factors in stage III non- small-cell lung cancer patients. Asian Pac J Cancer Prev. 2016. textithttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5454619/
[8] Roshan S and Dr. Rohini V. PREDICTION OF POST-SURGICAL SUR- VIVAL OF LUNG CANCER PATIENTS AFTER THORACIC SURGERY USING DATA MINING TECHNIQUES.2017
[9] Mark W. Onaitis, MD, Anthony P. Furnary, MD, Andrzej S. Kosinski, PhD, Sunghee Kim, PhD, Daniel Boffa, MD, Betty C. Tong, MD, Patricia Cowper, PhD, Jeffrey P. Jacobs, MD, Cameron D. Wright, MD, Joe B. Putnam, Jr, MD, and Felix G. Fernandez, MD, MSc.Prediction of Long- Term Survival After Lung Cancer Surgery for Elderly Patients in The Society of Thoracic Surgeons General Thoracic Surgery Database.
[10] https://www.techopedia.com/definition/14650/data-preprocessing
[11] V. Ilango, R. Subramanian,V. Vasudevan. A Five Step Procedure for Outlier Analysis in Data Mining.2012

