Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

Two Stage Classification Model to Classify Patients into Lower Variability Resource User Groups

M. Anitha[1], P. Sahithi[2], P. Lavanya[2], Sk. Amrin[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

ABSTRACT

Healthcare demand is growing in Australia and across the world. In Australia, the healthcare system comprises a mix of private and public organizations, such as hospitals, clinics, and aged care facilities. The Australian healthcare system is quite affordable and accessible because a large proportion of the expenditure, around 68%, is funded by the Australian government. The healthcare expenditure in 2015-16 was AUD 170.4 billion which was 10.0% of the GDP. Soaring healthcare costs and growing demand for services are increasing the pressure on the sustainability of the government-funded healthcare system. To be sustainable, we need to be more efficient in delivering healthcare services. We can schedule the care delivery process optimally and subsequently improve the efficiency of the system if demand for services is well known. However, there is a randomness in demand for services, and it is a cause of inefficiency in the healthcare delivery process.

Soaring healthcare costs and the growing demand for services require us to use healthcare resources more efficiently. Randomness in resource requirements makes the care delivery process less efficient. Our aim is to reduce the uncertainty in patients' resource requirements, and we achieve that objective by classifying patients into similar resource user groups. The conventional random forest., k-nearest neighbourhood (KNN) methods were resulted in poor classification, prediction performance.

In this work, we develop a two-stage classification model to classify patients into lower variability resource user groups by using electronic patient record. There are various statistical tools for classifying patients into lower variability resource user groups. However, classification and regression tree (CART) analysis is a more suitable method for analyzing healthcare data because it has some distinct features. For example, it can handle the interaction between predictor variables naturally, it is nonparametric in nature, and it is relatively insensitive to the curse of dimensionality.

Keywords: Classification and Regression tree, Length of stay patient, KNN.

1. INTRODUCTION

Healthcare demand is growing in Australia and across the world. In Australia, the healthcare system comprises a mix of private and public organizations, such as hospitals, clinics, and aged care facilities. The Australian healthcare system is quite affordable and accessible because a large proportion of the expenditure, around 68%, is funded by the Australian government [1]. The healthcare expenditure in 2015-16 was AUD 170.4 billion which was 10.0% of the GDP [1]. Soaring healthcare costs and growing demand for services are increasing the pressure on the sustainability of the government-funded healthcare system. To be sustainable, we need to be more efficient in delivering healthcare services. We can schedule the care delivery process optimally and subsequently improve the efficiency of the system if demand for services is well known. However, there is a randomness in demand for services, and it is a cause of inefficiency in the healthcare delivery process.

It is possible to design a deterministic system optimally to achieve a very high, ≥ 90%, utilization of the available resources. However, in a system with intrinsic randomness, improving the resource utilization diminishes the quality of services. For example, if we operate an intensive care unit (ICU) at a very high, ≥ 85%, occupancy level, we may need to refuse admissions frequently because of a

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

capacity shortage. To manage healthcare facilities efficiently, we need to minimize the effect of the randomness in demand for services on the efficiency of the system. The random arrival time and the uncertainty in resource requirements of each individual are the sources of variability in demand for services [2].

In hospitals, resources are bundled together, and medical professionals work in teams. A patient's resource consumption is measured by one's length of stay (LoS) at various care steps, such as the LoS in the ICU, the LoS in a surgical ward, and one's surgery duration. Therefore, the variability in resource requirements can be approximated by the variability in LoS. Moreover, in the case of elective operations, patients' arrival times are scheduled by the hospital administration. The remaining source of variability in the elective patient flow process is the randomness in LoS. We can manage a surgical suite more efficiently if we can predict patients' LoS accurately

2. LITERATURE SURVEY

McMullan et. al [3] described the patient demographic characteristics and organisational factors that influence length of stay (LOS) among emergency medical admissions. Also, to described differences in investigation practice among consultant physicians and to examine the impact of these on LOS. Faddy et. al [4] presented a relatively novel method for modeling length-of-stay data and assess the role of covariates, some of which are related to adverse events. To undertake critical comparisons with alternative models based on the gamma and log-normal distributions. To demonstrate the effect of poorly fitting models on decision-making. The model has the process of hospital stay organized into Markov phases/states that describe stay in hospital before discharge to an absorbing state. Admission is via state 1 and discharge from this first state would correspond to a short stay, with transitions to later states corresponding to longer stays. The resulting phase-type probability distributions provide a flexible modeling framework for length-of-stay data which are known to be awkward and difficult to fit to other distributions.

Liu et. al [5] evaluated the utility of adding "point of admission" automated laboratory and comorbidity measures—the Laboratory Acute Physiology Score (LAPS) and Comorbidity Point Score (COPS)—to risk adjustment models that are based on administrative data. They performed a retrospective analysis of 155,474 hospitalizations between 2002 and 2005 at 17 Northern California Kaiser Permanente hospitals. They evaluated the benefit of adding LAPS and COPS in linear regression models using full, trimmed, truncated, and log-transformed LOS, as well as in logistic and generalized linear models. Jiang et. al [6] Data mining techniques were applied to a classification task where various input variables were used to predict whether ALOS falls within normal category (less than or equal to mean plus three times standard deviation) or long category. Four models were built, and the ensemble model was selected as the best fit. Age and chronic disease were identified as the most important factors in predicting ALOS. Findings of this research can provide insights into where to start in the effort to improve the discharge rate and ultimately reduce costs for the hospital.

Freitas et. al [7] used hospital administrative data from inpatient episodes in public acute care hospitals in the Portuguese National Health Service (NHS), with discharges between years 2000 and 2009, together with some hospital characteristics. The dependent variable, LOS outliers, was calculated for each diagnosis related group (DRG) using a trim point defined for each year by the geometric mean plus two standard deviations. Hospitals were classified on the basis of administrative, economic, and teaching characteristics. They also studied the influence of comorbidities and readmissions. Logistic regression models, including a multivariable logistic regression, were used in the analysis. All the logistic regressions were fitted using generalized estimating equations (GEE). Carter and Potts et. al [8] investigated whether factors can be identified that significantly affect hospital length of stay from those available in an electronic patient record system, using primary total

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

knee replacements as an example. To investigated whether a model can be produced to predict the length of stay based on these factors to help resource planning and patient expectations on their length of stay. Data were extracted from the electronic patient record system for discharges from primary total knee operations from January 2007 to December 2011 (n = 2,130) at one UK hospital and analysed for their effect on length of stay using Mann-Whitney and Kruskal-Wallis tests for discrete data and Spearman's correlation coefficient for continuous data. Models for predicting length of stay for primary total knee replacements were tested using the Poisson regression and the negative binomial modelling techniques. Morton et. al [9] five supervised learning algorithms were applied to a subset of diabetic patient records from a large well known medical database. Using the Age Category, Indicator of Sex, Race, Expected Primary Payer, Admission Type, and APR-DRG variables we predicted short-term vs. long-term LOS for each patient, where short-term is defined as less than 3 days. They evaluated the performance of the MLR, SVM, SVM+, MTL, and RF supervised machine learning techniques using the AUC, ACC, and FS measures.

Sushmita et. al [10] used machine learning algorithms for accurate predictions of healthcare costs on publicly available claims and survey data. Specifically, they investigated the use of the regression trees, M5 model trees and random forest, to predict healthcare costs of individual patients given their prior medical (and cost) history. Overall, three observations showcase the utility of our research: (a) prior healthcare cost alone can be a good indicator for future healthcare cost, (b) M5 model tree technique led to very accurate future healthcare cost prediction, and (c) although state-of-the-art machine learning algorithms are also limited by skewed cost distributions in healthcare, for a large fraction (75%) of population, they were able to predict with higher accuracy using these algorithms. In particular, using M5 model trees we were able to accurately predict costs within less than $125 for 75% of the population when compared to prior techniques. Since models for predicting healthcare costs are often used to ascertain overall population health, this work is useful to evaluate future costs for large segments of disease populations with reasonably low error as demonstrated in our results on real-world publicly available datasets.

Al Taled et. al [11] introduces an approach for early prediction of LOS of stroke patients arriving at the Stroke Unit of King Fahad Bin Abdul-Aziz Hospital, Saudi Arabia. The approach involves a feature selection step based on information gain followed by a prediction model development step using different machine learning algorithms. Prediction results were compared in order to identify the best performing algorithm. Many experiments were performed with different settings.

Smith et. al [12] provides a detailed evaluation using simulations of the appropriateness of standard one-part generalized linear models (GLMs) compared to a recently developed marginalized two-part (MTP) model. The MTP model, unlike the one-part GLMs, explicitly accounts for the point mass at zero, yet takes the same form for the marginal mean as the commonly used GLM with log link, making the covariate effects directly comparable. They simulate data scenarios with varying sample sizes and percentages of zeros. One-part GLMs resulted in increased bias, lower than nominal coverage of confidence intervals, and inflated type I error rates, rendering them inappropriate for use with semicontinuous data. Even when distributional assumptions were violated, estimates of covariate effects and type I error rates under the MTP model remained robust.

Ting Zhu et. al [13] compares three models: a model combining seasonal regression and ARIMA, a multiplicative seasonal ARIMA (MSARIMA) model, and a combinatorial model based on MSARIMA and weighted Markov Chain models in generating forecasts of daily discharges. The models are applied to three years of discharge data of an entire hospital. Several performance measures like the direction of the symmetry value, normalized mean squared error, and mean absolute percentage error are utilized to capture the under- and overprediction in model selection. The findings

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

indicate that daily discharges can be forecast by using the proposed models. A number of important practical implications are discussed, such as the use of accurate forecasts in discharge planning, admission scheduling, and capacity reservation. Rouzbahman et. al [14] examined the extent to which analysis of clustered patient types can match predictions made by analyzing the entire dataset at once. After reviewing relevant literature, and explaining how data are summarized in each cluster of similar patients, they compare the results of predicting death, and length of stay (LOS) in the ICU1 using regression analysis on original and clustered data from the MIMIC II dataset. Clustering improved regression prediction accuracy for both death and LOS. They then show that clustering prior to regression also improved prediction of number of days to next emergency room visit for cancer patients. Thus, in all three prediction tasks that we investigated (involving two very different datasets), we found that clustering prior to regression analysis improved prediction accuracy.

Livieris et. al [15] presented a user-friendly decision support system for the prediction of hospitalized patients LoS which incorporates a two-level machine learning classifier. These numerical experiments revealed that the proposed classification technique exhibits better classification accuracy compared to some of the most popular and commonly used individual classification algorithms. Significant advantages of the presented software are the employment of a simple and user-friendly interface, its scalability due to its modular nature of design and implementation and its operating system neutrality. It is worth recalling that our expectation is that this work could be used as a reference for decision making in the admission process and strengthen the service system in hospitals by offering customized assistance according to patients' predicted hospitalization time.

3. PROPOSED SYSTEM

Initially, dataset is considered for implementing the overall system. The preprocessing operation is carried out to remove the different types of missing symbols, empty spaces. Then, two stage- CART algorithm is applied for prediction operation. Here, two stage CART is divided into DT-Classifier and DT- Regression. Finally, performance evaluation is carried out.

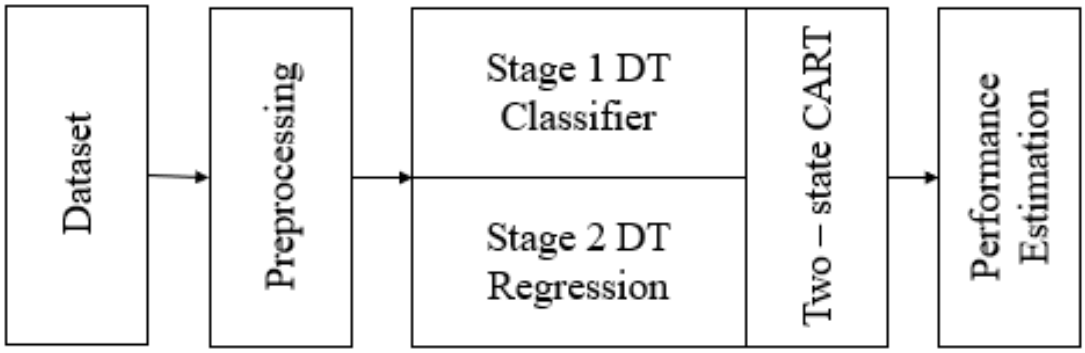


Fig. 1: Block diagram of proposed system.

Fig. 1 shows the block diagram of proposed system. In this work, we develop a two-stage classification model to classify patients into lower variability resource user groups by using electronic patient record. There are various statistical tools for classifying patients into lower variability resource user groups. However, classification and regression tree (CART) analysis is a more suitable method for analyzing healthcare data because it has some distinct features. For example, it can handle the interaction between predictor variables naturally, it is nonparametric in nature, and it is relatively insensitive to the curse of dimensionality.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

3.1 Dataset

LOS_model: A simulated hospital length-of-stay dataset

This vignette details why the LOS_model dataset was created, how to load it, and gives examples of use for learning/teaching regression modelling using Generalized Linear Models (GLMs), and related techniques.

The data were created specifically for regression tutorials, simulating a small set of data on hospital in-patient spells. The data are 10 sets of 30 simulated patient records, representing 10 different hospitals ("Trusts"). The dataset contains

- ID: an integer value patient number
- Organisation: A factor, containing hospital name, e.g. "Trust1"
- Age: an integer representing patient age in years
- LOS: 'Length of Stay,' an integer representing the number of days a patient was in hospital
- Death: an integer flag (0 or 1) representing whether a patient died

3.2 Preprocessing

Data Preprocessing in Machine learning

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

Why do we need Data Preprocessing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

2) Importing Libraries

In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:

Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

import numpy as nm

Here we have used nm, which is a short name for Numpy, and it will be used in the whole program.

Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code. It will be imported as below:

import matplotlib.pyplot as mpt

Here we have used mpt as a short name for this library.

Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library. It will be imported as below:

Here, we have used pd as a short name for this library. Consider the below image:

```
1 # importing libraries
2 import numpy as nm
3 import matplotlib.pyplot as mtp
4 import pandas as pd
5
```

3) Importing the Datasets

Now we need to import the datasets which we have collected for our machine learning project. But before importing a dataset, we need to set the current directory as a working directory. To set a working directory in Spyder IDE, we need to follow the below steps:

Save your Python file in the directory which contains dataset.

Go to File explorer option in Spyder IDE and select the required directory.

Click on F5 button or run option to execute the file.

4) Handling Missing data

The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

Ways to handle missing data:

There are mainly two ways to handle missing data, which are:

By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.

By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc.

5) Encoding Categorical data

Categorical data is data which has some categories such as, in our dataset; there are two categorical variables, Country, and Purchased.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

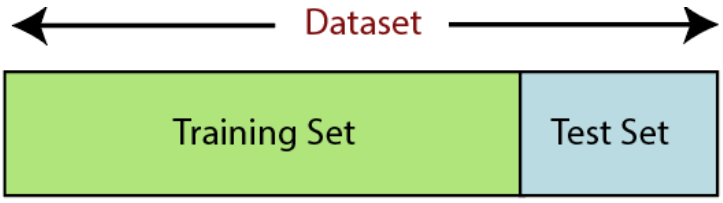
Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So, it is necessary to encode these categorical variables into numbers.

6) Splitting the Dataset into the Training set and Test set

In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:



Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

For splitting the dataset, we will use the below lines of code:

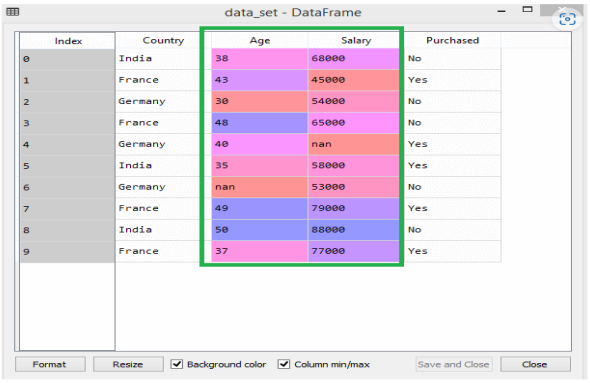
```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.2, random_state=0)
```

7) Feature Scaling

Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no variable dominate the other variable.

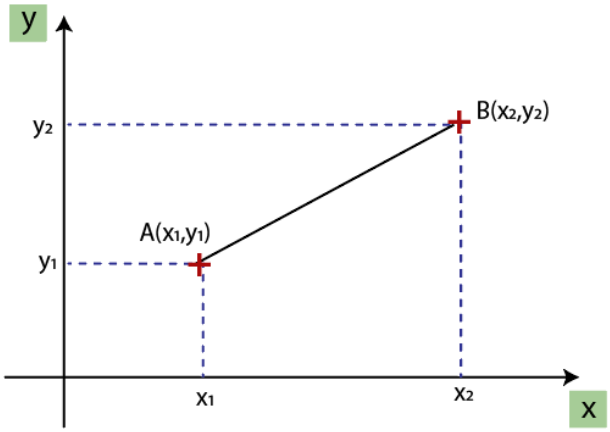
Consider the below dataset:



ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

As we can see, the age and salary column values are not on the same scale. A machine learning model is based on Euclidean distance, and if we do not scale the variable, then it will cause some issue in our machine learning model.

Euclidean distance is given as:

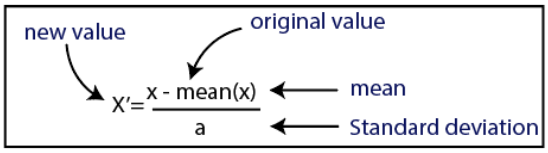


$$\text{Euclidean Distance Between A and B} = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

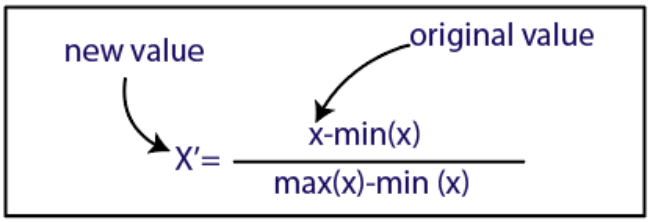
If we compute any two values from age and salary, then salary values will dominate the age values, and it will produce an incorrect result. So to remove this issue, we need to perform feature scaling for machine learning.

There are two ways to perform feature scaling in machine learning:

Standardization



Normalization



3.3 CART Algorithm

There are various statistical techniques to cluster and classify data. Most of the clustering techniques come under unsupervised learning techniques. In unsupervised learning, the model is not trained with known results, and the algorithm clusters data points into groups according to their statistical properties, such as the mean, the median, and the variance. However, under supervised learning techniques, the predictions are based on the training sample containing joint observations of dependent and independent variables. Statistical techniques such as multivariate regression analysis, logistic regression analysis, Fisher's discriminant analysis, support vector machine, neural networks, and decision tree analysis or classification and regression tree (CART) analysis, are some of the commonly used classification techniques.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

In CART analysis, recursively partition the observations into progressively smaller groups. Each partition is a binary split based on single independent variable. Let us assume that our data have k patient attributes which are independent variables, and a response variable, say LoS, for each of N observations, that is, (A_i, LoS_i) for i = 1, 2, ..N, with $A_i = (A_{i1}, A_{i2}, .., A_{ik})$. Now the algorithm will select attribute A_{ij} , and split the region R (set of observations in a partition) into two regions or partitions, R_1 and R_2, in such a way, so that,

$$\min_{j,s}\left[\sum_{A_i\in R_1(j,s)}(LoS_i - c_1)^2 + \sum_{A_i\in R_2(j,s)}(LoS_i - c_2)^2\right]$$

Where $\hat{c}_1 = ave(LoS_i|A_i \in R_1(j,s))$, and $\hat{c}_2 = ave(LoS_i|A_i \in R_2(j,s))$. In the case of a categorical attribute Aij , the point s will be a subset of values of A_{ij}, and the two regions will be $A_I \in R_1(j,s)$ and $A_I \in R_1(j,s')$

Where

$$s \cap s' 0 = \varphi.$$

Now, the two groups become the intermediate nodes, and the process is repeated recursively for each intermediate node until a stopping criterion is met. However, the resultant tree could be very large. The tree size is an important tuning parameter that decides model complexity. A very large tree may overfit the data, whereas a very small tree may ignore some of the natural clusters. To overcome this difficulty, we generally grow a maximum size tree and then prune it to an optimal tree by using some cost complexity criterion. If the leaf nodes are indexed by m such that node m represents partition R_m, then for a tree T, we define,

$$N_m = \#\{A_i \in R_m\},$$
$$\hat{c}_m = \frac{1}{N_m}\sum_{A_i\in R_m} LoS_i,$$
$$Q_m(T) = \frac{1}{N_m}\sum_{A_i\in R_m}(LoS_i - \hat{c}_m)^2,$$

and the cost complexity criterion is defined as,

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|,$$

where Q_m is the average unexplained variability at node m, and $C\alpha$ is the cost function with penalty factor α. The parameter α determines the trade-off between the goodness of fit of the tree and its size. As the value of α increases, the tree size decreases. In our analysis, we used a ten-fold cross-validation (CV), that is, a CV where we split the data randomly into ten parts and used nine parts of the data as training data and one part as validation data. We repeat the process ten times with each of the ten parts used exactly once as validation data, and we use the average CV error across all ten trials to select the final tree. Since the average CV error is a random variable with a significant amount of variability in it, its value for an over-fitted tree can be lower than that for an optimal tree. Therefore, to avoid over-fitting, we selected α that corresponded to the simplest tree, after which there was no significant reduction in the average CV error.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

Let us say that our optimal tree contains m partitions after pruning, then we can calculate the reduction in total variance as follows. If

$$\hat{c} = \frac{1}{N}\sum_{i=1}^{N} LoS_i,$$
$$V_T = \frac{1}{N}\sum_{i=1}^{N}(LoS_i - c)^2,$$
$$V_R = \frac{1}{N}\sum_{A_i \in R_m}(LoS_i - c_m)^2,$$

Then

$$V_E(in\%) = \frac{V_T - V_R}{V_T} \times 100,$$

where V_E is the variance explained by the fitter tree model, $V_T$is the total variance in the data, and V_R is the residual variance or unexplained variance. The ratio of V_R to V_T is defined as the relative error, and the ratio of average CV to V_T is defined as the relative CV error. The lower the relative CV error, the better the fitted model classifies new data. We use relative CV error to compare the performance of the CART and other methods.

We used CART analysis to classify patients into groups because of the following reasons.

- It is very simple and intuitive. The results from CART analysis is basically the partitioning of patients into groups according to their attributes. This partitioning may easily be verified by medical professionals and approved for implementation.
- It is a non-parametric method and is a suitable technique for mixed datasets. Patient attributes are generally of a mixed type. For example, gender is a categorical variable, age is an ordinal variable, and LoS is a continuous variable. CART analysis can handle all type of variables without any difficulty. Moreover, there are no assumptions regarding the distribution of the variables or the residual error terms.
- It does not suffer from the curse of dimensionality. In statistics, the curse of dimensionality is related to the fact that a classifier loses its credibility when it is defined on a high-dimensional space. In healthcare data, there are a large number of patient attributes which might be capable of explaining the variability in LoS. In general, it is difficult for us to select the most relevant attributes. If we fit a classifier based on all the attributes, then the model may suffer from the curse of dimensionality. However, because CART analysis only considers one variable at a time while making a partitioning rule, it does not suffer from the curse of dimensionality
- It can model complex interactions between independent variables naturally without increasing model complexity. In the CART analysis, the same variable can be used in different parts of the tree to make a split. However, in other classification techniques, we need to add one additional variable corresponding to each interaction term which will further increase the dimension of the search space.
- If some predictor variables are missing in the data, the CART analysis uses the next most competitive predictor as a surrogate to make a binary split. Therefore, we do not need to remove data entries with missing fields.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

In order to evaluate the performance of the CART, we compare the CART analysis results with that of the random forest (RF) and the k-nearest neighbor (knn) regression. In regression tree model, we repeatedly make binary splits in the response variable according to the most promising predictor variable at each node until the stopping criteria are met. In RF, we grow a large number of uncorrelated classification or regression trees, and we compute the response variable by averaging predictions across all the trees. In knn, we select k training points closest to the new data, and we compute the response variable of the new data by averaging the response values among k nearest neighbors. Despite its simplicity, knn regression is a very competitive method particularly for the data where each class has various possible prototypes, and the decision boundary is very irregular.

Statistical Model

There are two possible ways to classify data into groups by using CART analysis. These are classification tree analysis and regression tree analysis. In classification tree analysis, we first need to cluster data points into the required number of groups and then a classification tree model is fitted to training data. By using the fitted classification model, we can predict the cluster of new data. Although there are several statistical methods available to cluster data points into groups, most of these methods come under unsupervised learning techniques. This means that the clusters are obtained based solely on some statistical properties of the LoS and is irrelevant to the values of other covariates which is clinically not sensible. If we use this grouping as an independent variable to figure out the covariates that are capable of explaining the variability, our results will be sensitive to the pre-processing. As a result, our classification model may not fit well for the assigned classes.

In the second method, we split the data according to each possible partition of each independent variable, and then we greedily select the split that minimises the weighted sum of variation in the dependent variable. This method is more advantageous for our problem as it clusters patients according to their LoS and their clinically crucial attributes. When we clustered patients using the regression tree method, we observed that there was some overlap between different LoS groups.

3.4 Decision Tree with CART Algorithm

Decision tree is one of most basic machine learning algorithms which has wide array of use cases which is easy to interpret & implement. We can use decision tree for both regression & classification tasks. In this article we will try to understand the basics of Decision Tree algorithm. Then how Decision tree gets generated from the training data set using CART algorithm.

About Decision Tree

Decision tree is a non-parametric supervised learning technique, it is a tree of multiple decision rules, all these rules will be derived from the data features. It is one of most easy to understand & explainable machine learning algorithm. This ML algorithm is the most fundamental components of Random Forest, which are most popular & powerful ML algorithm.

Structure of Decision Tree

In the below image I tried to show how a decision tree would look like. Each internal node represents a segment or region. With respect to tree analogy, segments or regions are nodes or leaves of the tree.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

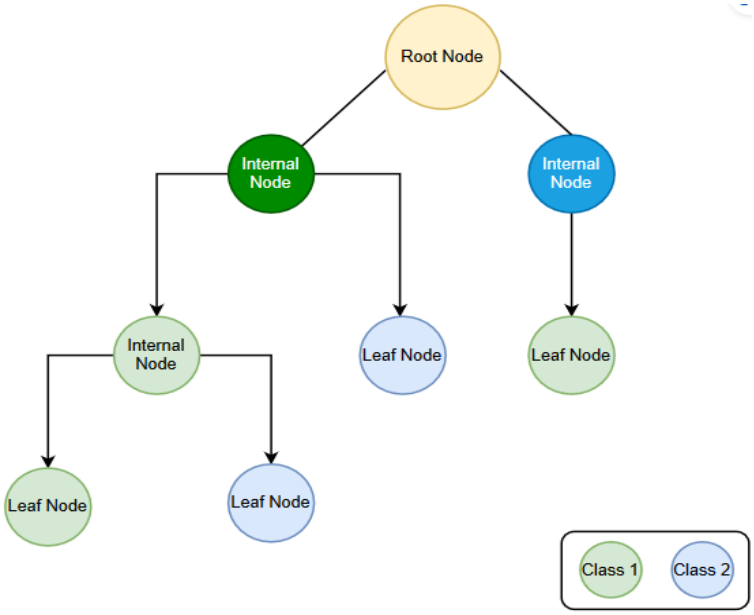


Fig. 2: Decision tree structure.

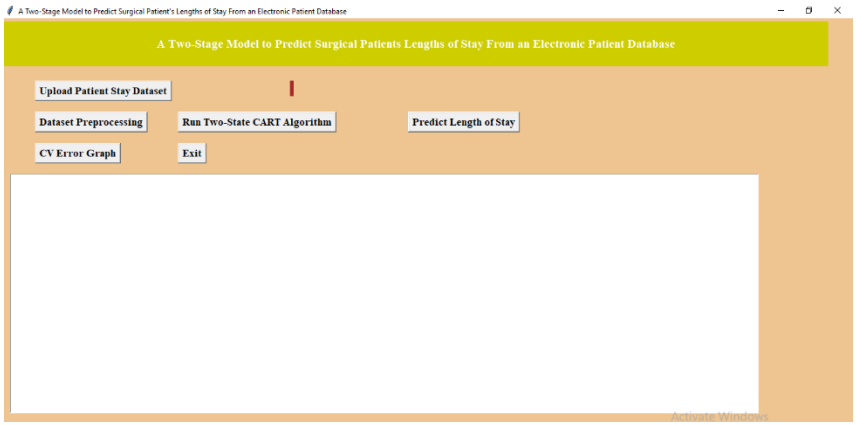
Root Node: This is the first node which is our training data set.

Internal Node: This is the point where subgroup is split to a new sub-group or leaf node. We can call this as a decision node as well because this is where node splits further based on the best attribute of your sub-group.

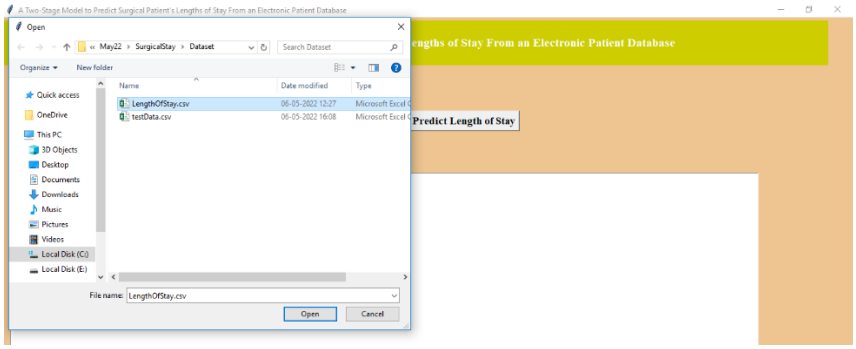
Leaf Node: Final node from any internal node, this holds the decision.

4. RESULTS AND DISCUSSION

To run project double click on 'run.bat' file to get below output.

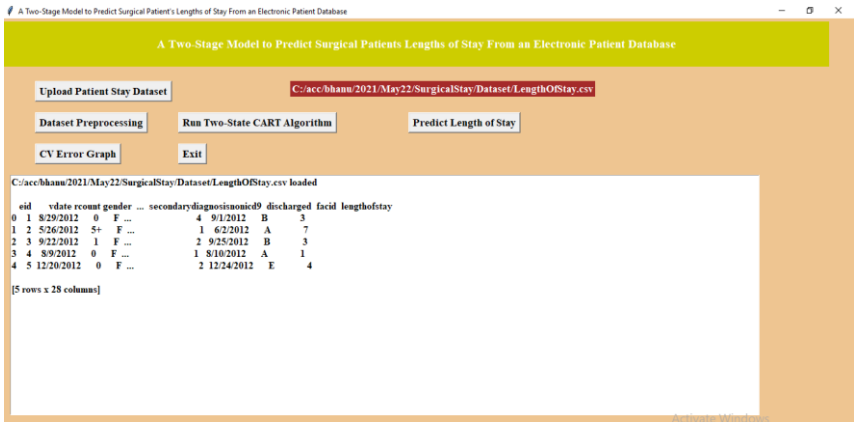


In above screen click on 'Upload Patient Stay Dataset' button to upload dataset

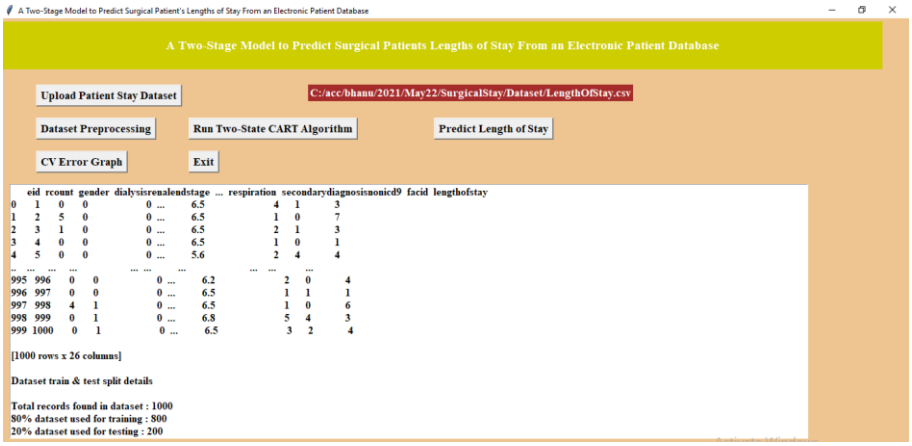


ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

In above screen selecting and uploading dataset and then click on 'Open' button to get below output



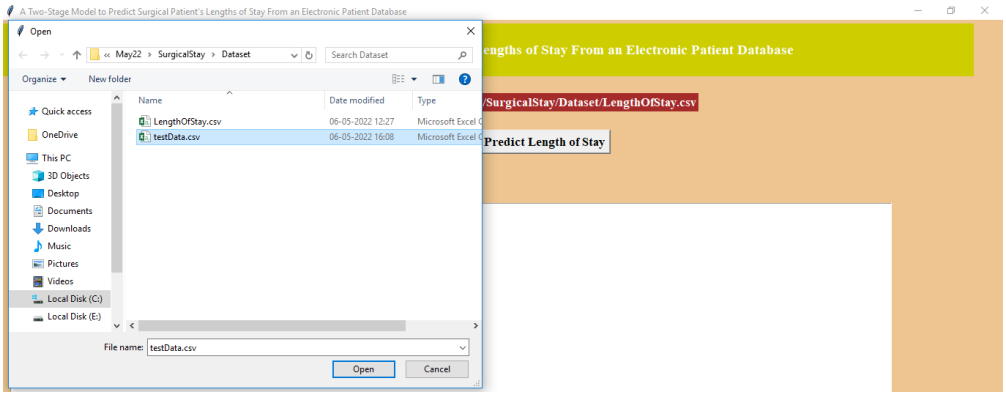
In above screen dataset loaded and we can see dataset contains some non-numeric values so we need to process dataset to convert or encode non-numeric to numeric values.



In above screen we can see all dataset values converted to numeric and we can see dataset contains 1000 records and 800 using for training and 200 for testing.

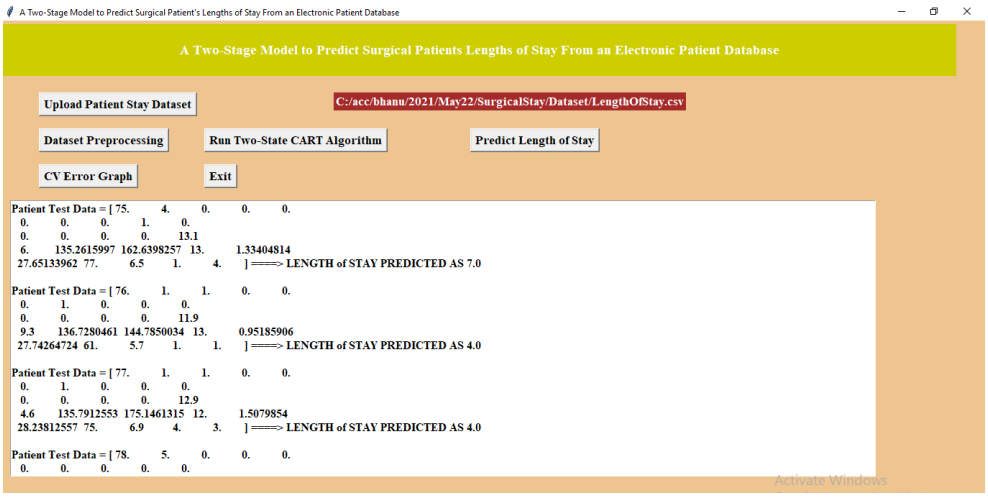
1000	KNN	0.906	0.906
1000	Two-Stage CART	0.5730000000000001	0.55

In above screen we can see misclassification error rate for each algorithm in stage 1 and 2 and in above screen we can see proposed two stage CART error rate is less compare to KNN.



In above screen selecting and uploading 'testData.csv' file and then click on 'Open' button to load test data and get below prediction output

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023



In above screen in square bracket you can see the patient test data and after arrow symbol you can see length of stay as 7 days and similarly for each test record you can see predicted length of days.

5. CONCLUSION

In this work, we developed a novel method to classify patients into lower variability LoS groups by using the electronic patient database. We deployed machine learning techniques to cluster patients with similar resource requirements which is useful for tactical and operational planning. The results show that the CART analysis is a useful tool for clustering patients, and it can perform feature selection even when there are many predictor variables. We can fit the probability distributions to each partition obtained from the CART analysis which is useful for developing a robust tactical and operational planning policies. By using the novel approach, we have developed, we were able to reduce the relative CV error in predictions further up to 9.0%. We compared the performance of the CART with that of the k-nearest neighbor regression (knn) and the decision tree. The results of the knn regression were quite competitive, but less informative. The decision tree provided slightly better predictions regarding the remaining variability.

However, the obtained predictions were average values of sampled scenarios, and not the accurate LoS scenario realizations. Therefore, we recommend using the stage-II clustering scheme for strategic planning, and the RF classification model presented in subsection V-D for predicting new patients' classes. For the future work, we would apply the CART method to cluster patients with similar resource requirements in other healthcare facilities, such as the operating theater, and the ICU, and we will investigate the results. We would also explore some models to predict the readmission rate and the risk of the ICU admission after operation by using the CART analysis or other tree-based methods. To improve the accuracy of the CART algorithm, which is a greedy one, we would explore some optimization techniques. While performing partitioning recursively, it sequences partitioning greedily instead of finding the optimal partitioning sequence. As a result, the algorithm may provide locally optimal partitions as a final partition. Since the covariates attributable to the variability in LoS are categorical, the problem may be modeled as a combinatorial optimization problem to obtain an optimal partitioning. In future, we will investigate some of these modeling techniques to solve this problem

REFERENCES

[1] AIHW, "Australia's health series no. 15. Cat. no. AUS 199," Canberra: AIHW, 2015, "[Online; accessed 25-September-2022]". [Online]. Available:

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 07, 2023

https://www.aihw.gov.au/reports/australias-health/ australias-health-2016/contents/summary

[2] P. R. Harper, "A framework for operational modelling of hospital resources," Health Care Manag. Sci., vol. 5, no. 3, pp. 165–173, 2002.

[3] R. McMullan, B. Silke, K. Bennett and S.Callachand. "Resource utilisation, length of hospital stay, and pattern of investigation during acute medical hospital admission". Postgrad Med J. 2004 Jan;80(939):23-6. doi: 10.1136/pmj.2003.007500. PMID: 14760174; PMCID: PMC1757957.

[4] M. Faddy, N. Graves, and A. Pettitt, "Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and lognormal distributions", Value Heal., vol. 12, no. 2, pp. 309–314, 2009

[5] V. Liu, P. Kipnis, M. K. Gould and G. J. Escobar. "Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables". Med Care. 2010;48(8):739–44. https://doi.org/10.1097/MLR.0b013e3181e359f3

[6] X. Jiang, X. Qu and L. Davis. "Using Data Mining to Analyze Patient Discharge Data for an Urban Hospital". In: DMIN; 2010. p. 139–144.

[7] A. Freitas, T. Silva-Costa and F. Lopes. "Factors influencing hospital high length of stay outliers". BMC Health Serv Res 12, 265 (2012). https://doi.org/10.1186/1472-6963-12-265

[8] E. M. Carter and H. W. Potts. "Predicting length of stay from an electronic patient record system: a primary total knee replacement example". BMC Med Inform Decis Mak 14, 26 (2014). https://doi.org/10.1186/1472-6947-14-26

[9] A. Morton, E. Marzban, G. Giannoulis, A. Patel, R. Aparasu and I. A. Kakadiaris, "A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay among Diabetic Patients", 2014 13th International Conference on Machine Learning and Applications, 2014, pp. 428-431, doi: 10.1109/ICMLA.2014.76.

[10] S. Sushmita, S. Newman and J. Marquardt. "Population cost prediction on public healthcare datasets". In: Proceedings of the 5th international conference on digital health 2015. ACM; 2015. Pp. 87–94. https://doi.org/10.1145/2750511.2750521

[11] A. R. Al Taleb, M. Hoque, A. Hasanat and M. B. Khan, "Application of data mining techniques to predict length of stay of stroke patients", 2017 International Conference on Informatics, Health & Technology (ICIHT), 2017, pp. 1-5, doi: 10.1109/ICIHT.2017.7899004.

[12] V. A. Smith, B. Neelon and M. L. Maciejewski. "Two parts are better than one: modeling marginal means of semicontinuous data", Health Serv Outcomes Res Method 17, 198–218 (2017). https://doi.org/10.1007/s10742-017-0169-9

[13] Ting Zhu, Li Luo, Xinli Zhang, Yingkang Shi and Wenwu Shen. "Time-Series Approaches for Forecasting the Number of Hospital Daily Discharged Inpatients", IEEE J Biomed Health Inform. 2017 Mar;21(2):515-526. doi: 10.1109/JBHI.2015.2511820. Epub 2015 Dec 23. PMID: 28055928.

[14] M. Rouzbahman, A. Jovicic and M. Chignell, "Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU?", in IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 3, pp. 851-858, May 2017, doi: 10.1109/JBHI.2016.2525731.

[15] I. E. Livieris, T. Kotsilieris, I. Dimopoulos and P. Pintelas. "Decision Support Software for Forecasting Patient's Length of Stay", Algorithms 2018, 11, 199. https://doi.org/10.3390/a11120199