EMITTER International Journal of Engineering Technology
Vol. 12, No. 2, December 2024, pp. 128~149
DOI : 10.24003/emitter.v12i2.835

Impact of Principal Component Analysis on the Performance of Machine Learning Models for the Prediction of Length of Stay of Patients

Jagriti Gupta[1], Naresh Sharma[1], Sandeep Aggarwal[2]

[1]School of Engineering and Sciences, G D Goenka University, Gurugram, India
[2]Department of Management Science, MDU-CPAS, MDU-Rohtak, India
Corresponding author: naresh.sharma@gdgu.org

Received April 19, 2024; Revised October 5, 2024; Accepted November 26, 2024

Abstract

Patient inflow, limited resources, criticality of diseases and service quality factors have made it essential for the hospital administration to predict the length of stay (LOS) for inpatients as well as outpatients. An efficient and effective LOS prediction tool can improve the patient care and minimize the cost of service by increasing the efficiency of the system through optimal allocation of available resources in the hospital. For predicting patient's LOS, machine learning (ML) models can have encouraging results. In this paper, five ML algorithms, namely linear regression, k- nearest neighbours, decision trees, random forest, and gradient boosting regression, have been used to predict the LOS for the patients admitted to the hospital with some medical history, laboratory measurements, and vital signs collected before admission. Additionally, the impact of principal component analysis (PCA) has been analyzed on the predictive performance of all ML algorithms. A five-fold cross-validation technique has been used to validate the results of proposed ML model. The results concluded that the RF and GB model performs better with R^2 score of 0.856 and 0.855 respectively among all the ML models without using PCA. However, the accuracy of all the models increased with the PCA except KNN and LR. The GB model when used with principal components has R^2 score and MSE approximate to 0.908 and 0.49 respectively compared to the model that incorporates with the original data. Additionally, PCA has an advantageous effect on the DT, RF and GB models. Therefore, LOS for new patients can be predicted effectively using the proposed tree-based RF and GB model with using PCA.

Keywords: machine learning models, length of stay prediction, regression, principal component analysis

1. INTRODUCTION

The number of hospitalization each year is expected to rise by 42% by 2050, from 4.66 million to 6.72 million days. [1]. This growth would result in much greater healthcare costs for each country. So, to keep up with the rising cost and demand for hospitals in 2050, there needs to be better planning,

Copyright © 2024 EMITTER International Journal of Engineering Technology - Published by PENS

allocation of beds, and admission.[2]. The inflow of inpatients and outpatients into hospital units has increased with the rise of medical issues in recent years. This has put the hospital administration under constant pressure to manage costs and services. Increased demand for quality services and intensive care with limited resources has forced the medical teams to predict the accurate flow of patients in different units along with length of stay (LOS) for the hospitalized patients. The time spent in the hospital from admission to the process of being discharged is defined as the patient's length of stay [3].

An estimate of the patient's LOS in the hospital helps in better planning for bed allocation, prescribing consultants for patients with multiple disorders, and planning the date of discharge for elderly patients [4]. An accurate estimation of discharge time of an admitted patient can help the management team allow more admissions for treatment in the hospital [5]. Indicators of patient waiting time and LOS are of great help in the emergency department, and such indicators of delayed patient care [6] can help in achieving a better health care system. It was found by Zhuang et al. [7] that increased death rates during COVID-19 can be attributed to non-availability of beds and other resources in the hospital. These resources can be optimised to an extent by proper estimation of LOS for different indoor patients using the ML models with higher accuracy. An accurate estimation of LOS will not only improve services but also decrease the readmission rate.

Researchers have used statistical methods for predicting the LOS in the past [8]. However, the availability of large data sets and demand for more accurate results have motivated the researcher to use machine learning algorithms (ML) for predicting waiting time and LOS of different types of patients at different hospitals. Lequertier et al. [9] found that modern systems of healthcare management generate a large amount of data related to patient's medical histories, symptoms, lab results, departments, medical costs, availability of beds, and several types of diseases. Various ML models and statistical techniques can be applied to this kind of data available in the databases to facilitate the allocation of internal resources, reduce service time and minimise costs. This will help in achieving the balance between cost of quality services, reducing waiting times, reducing costs for patients, and enhancing the overall experience of hospital management [10].

The quest for quality services by hospital management has motivated the researcher to use different ML models to accurately predict the waiting time for various stages in the hospitals. Different departments can be modeled independently for different processes, units, and stages such as radiology, pathology, ICU unit, emergency department (ED) and others. Apart from the criticality of the diseases and the type of services, LOS can also be modelled with demographic factors such as age, marital status, and employment for better insight [11].

However, the previous literature lacks a comprehensive analysis of the impact of dimensional reduction techniques, such as principal component analysis, on the performance of ML models in the context of LOS prediction.

Addressing this gap is crucial for developing more accurate and efficient models that can enhance hospital management, resource allocation, and patient care.

In this paper, researchers have developed a model for predicting LOS of patients prior to admission using ML algorithms. Significant variables influencing the patient's LOS during admission have been identified through literature and in consultation with the practicing doctors. Five ML algorithms including linear regression (LR) [12], k- nearest neighbors (KNN) [13], decision tree (DT) [14], random forest (RF) [15], and gradient boosting (GB) have been used to develop predictive models for the selected datasets. To overcome the computational complexity of higher dimensional datasets and increase the accuracy of proposed predictive model, a principal component analysis (PCA) which is the dimensional reduction approach has been used. Furthermore, the performance of proposed ML models is thus compared with and without PCA to achieve the objective. Evaluation metrics such as R^2-score and MSE for each model have been used to determine the impact of PCA on predictive power of algorithms.

2. RELATED WORKS

In recent years, the application of ML has broadly increased for predicting future outcomes in various sectors, including the healthcare industry. Historical data is used in healthcare sectors to predict the inflow of patients, waiting times for different service units, LOS for patients admitted to hospitals, and even the disease a patient might be suffering from. For predicting LOS at the time of admission, previous researchers have mainly used ML algorithms like linear regression (LR), decision tree (DT), artificial neural network (ANN), random forest (RF), gradient boosting (GB) and k- nearest neighbors (KNN) to predict the LOS in various departments of hospitals [24, 25].

Gentimis et al. [28] predicted the LOS at the point where patients are transferred out of the ICU unit by using a neural network (NN) and RF model based on the parameters of admission datasets. The neural network predictive model performed better with an accuracy of approximately 80%. Morton et al.. [4] predicted the LOS for the diabetic patient using multitask, multiple linear regression (MLR) and support vector machine (SVM) algorithms. Demographic information (age, gender), hospital related information (location, number of beds), type of admission, severity measures, number of diagnoses, total charges of hospital, and LOS were the major attributes considered for modelling the result. SVM performed better in prediction with an accuracy of 68% compared to other algorithms.

Hijry & Olawoyin [29] built a model to predict the LOS of patients in the emergency department of a public hospital in Mecca (Saudi Arabia) by creating an integrated model with the help of an artificial neural network (ANN), linear regression and logistic regression. ANN performed better than other algorithms with an accuracy of 78.29%. Age, gender, number of patients, and diagnostic category were found to be the most important features affecting the

result of model. A review paper [30] shows that performance of the different models used by researchers in past studies varied significantly with the different input features and outcome metrics. Results obtained from ML algorithms have significantly outperformed estimates made using statistical methods.

Naemi et al. [31] predicted the LOS of patients upon admission to the emergency department. They addressed the effect of data skewness and missing values in the datasets on the performance of ML models using regression and classification techniques. The ML models, including SVM, RF, neural networks, extreme gradient boosting, and DT were implemented for predictive modelling. The multivariant imputation technique was used for filling missing values and the SMOTE technique was applied to balance the datasets. It was found that by addressing these challenges, accuracy of model on average increased by 32%. RF and NN models performed better among all algorithms.

Siddiqa et al. [32] used multiple ML algorithms to build a model for predicting the LOS of patients using a datasets obtained from New York Hospital. Six different models were used to predict the performances using the scores of MSE and R^2. It was concluded that with an MSE score of 5% and an R^2 score of 92%, RF model outperformed the other models namely multiple linear regression, lasso regression, Ridge regression, DT, and extreme gradient boosting (XGBoost). This study can be of significant help for predicting admission days in critical services like surgery, ICU, and cardiac arrest, as these departments demand more funds from the hospital and use more resources.

Aghajani & Kargari [33] determined that the type of surgery, average visits per day, and hospitalisation days before surgery influence the LOS for surgery department. Among the three algorithms KNN, naïve bayes (NB), and DT which are used to predict the LOS in the hospital during admission, the DT performed better than others with an accuracy of 88.29%. López-cheda et al. [34] predicted the LOS and the time to discharge of patients in ICU from the hospital using a nonparametric model. The datasets used in this study includes 10454 confirmed cases of COVID-19 reported in Galicia. The data resulted in an average stay of 11 days and found that LOS is different for female and male patients. They also found that age and gender features affect the LOS.

Y. Chen [35] analyzed the hospitalization data using an improvised nonlinear weighted XGBoost technique. The improved XGBoost algorithm performed with 82% accuracy and predicted number of beds available in a reliable manner compared to DT, NB, and K-NN. The multiple ML regression algorithms were compared to predict the LOS of different types of diseased patients [36]. They developed an ML model by using an open-source datasets and found the most prominent features that can affect the LOS of patients. With the lowest MAE of 0.44 and highest R^2 of 0.94, GB model performed best. Adawiyah et al. [37] created a system to predict LOS of patients in hospital using NN by using 3055 observations. They used NN with default parameters, with hyperparameter optimization techniques such as grid search and random

search and found that grid search method gives the highest performance with accuracy 94.7%.

Wan et al. [38] developed four ML models such as LR, SVR, ANN, and XGBoost to predict the overall strength of concrete depending on its mixture composition. Each model is implemented on three datasets, such as a datasets with 8 original features, a datasets with 6 principal components and a datasets with 6 manually chosen features. The XGBoost model performed best with manually selected features by obtaining highest R^2 score 0.93, while SVR performed best with PCA selected features with R^2 score of 0.91. Additionally, they found that the PCA technique negatively impacts the performance of ANN and LR models. Gupta et al. [39] proposed ML models such as RF and ANN to predict the disease in the Parkinson's datasets. They used PCA as a dimensionality reduction technique and applied it before implementing the model. They found that the ANN model outperformed with an accuracy of 97% when PCA was applied, while accuracy of RF model decreased with PCA from 89% to 79%.

From the review of available literature, it has been observed that most of the research work is based on a small set of datasets of patients containing features associated with one particular disease or with two, to calculate the LOS of patients. To test the prediction accuracy of the ML models, we selected a dataset with a large number of features, including medical history, lab results, vital signs, and readmission rate. However, there are currently no in-depth studies examining the impact of dimensional reduction techniques on the performance of different ML models for this application. In this paper, we propose PCA-based ML models for predicting the LOS of patients prior to admission to the hospital. Also, performance of ML models has been compared with and without PCA.

3. ORIGINALITY

The originality of this paper lies in the application of PCA technique in predicting LOS of patients in general hospital. The prediction of LOS of patients at the time of admission can help the hospital staff to manage resources efficiently. Before preprocessing, the datasets consist of 28 columns with numerical and categorical features. This dataset includes the demographic information, medical history of patients, vital signs and LOS of patients in days. The regression ML model is developed to predict the LOS of patients. Due to large number of features, there is complex model developed so there is need to reduce features to reduce complexity of the model. PCA is applied to reduce the dimension of the datasets which can increase the accuracy of model.

4. SYSTEM DESIGN

In this paper, we predict the length of stay of patients at admission by using a dataset consisting of the patient's medical history, lab test results, gender, and readmission rate within 180 days. The proposed model of the study is shown in Figure 1. Firstly, the datasets have been pre-processed before

fitting the ML models. In the second stage, the pre-processed datasets were split into a training set and a testing set. Afterward, feature extraction technique (PCA) was applied to both datasets to reduce the dimension of data and finally, selected ML algorithms such as LR, KNN, DT, RF, and GB with five-fold cross validation were trained on reduced training datasets, and models were evaluated on a test set to evaluate and compare the results. All this work was done in Jupyter Lab using the Python programming language. A brief outline of these ML algorithms and PCA is given below**.**

4.1.1 Linear Regression (LR)

Linear regression [12] is one of the essential regression algorithm for determining the relationship between a single continuous dependent variable and a number of independent variables. The prediction results can be derived from the equation (1)

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} \dots \dots \dots . b_p x_{ip} \quad (1)$$

where, y represents the output value, b is a constant coefficient, and x represents input variables. It works on the principle of mean squared error (MSE)[16].

4.1.2 K-Nearest Neighbors (KNN)

The KNN algorithm [13] is used for both regression and classification tasks. This is a non-parametric algorithm that makes no assumptions about the underlying data. k represents the number of nearest neighbors from the query point. This algorithm runs several times with different values of k, and chooses the optimal value of k that reduces the error to give a prediction with better accuracy. To find the distance between the nearest data points and the query point, Euclidean distance formula can be seen in equation (2).

$$\mathrm{d(x, y)} = \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{y_i} - \mathrm{x_i})^2} \quad (2)$$

The closest data points were determined based on their distances from the query point. In case of regression problem, we estimate the value of the dependent variable by averaging the values of its k nearest neighbours [17].

4.1.3 Decision Tree (DT)

This ML algorithm is used for both classification and regression problems [14]. It is quite easy to understand the results of decision tree as this algorithm works on the splitting criteria. The datasets were recursively split into subsets based on different features and their respective thresholds. variable at a time. The feature for each split is determined by finding the split that minimizes the mean squared error. To predict the target variable for a new instance, the algorithm follows splitting decisions based on the feature values until it

reaches the leaf node [18]. The value predicted at each leaf node is then chosen as the output.

4.1.4 Random Forest (RF)

This algorithm is a collection of multiple decision trees constructed during training [15]. This algorithm randomly selects a subset from original datasets to create multiple bootstrap samples and selects a subset of features from original datasets. Each bootstrap sample and subset of features are used to train the decision tree separately [19]. After training, the test instance passes through all decision trees and makes predictions for each tree. In a regression problem, RF predicts the target instance by taking the average of predictions from all decision trees. The prediction of RF model is calculated using the equation (4)

$$\mathbf{y} = \frac{\mathbf{1}}{\mathbf{n}}\left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i}\right) \tag{3}$$

where, $\boldsymbol{y}$ is the predicted value of given test instance, $\boldsymbol{n}$ is the number of decision trees in random forest model, and $\boldsymbol{y_i}$ is the predicted value of $\boldsymbol{the\ ith}$ decision tree.

4.1.5 Gradient Boosting (GB)

The GB algorithm is the ensemble ML learning algorithm used for regression and classification tasks [20]. It builds sequentially by reducing the error of previous models (weak predictors). In this algorithm, each decision tree is built one after the other to improve the deficiencies of previous one. This algorithm aggregates the results of decision trees during the process itself.

4.2 Principal Component Analysis (PCA)

In the feature extraction technique, new features are constructed from existing ones based on linear and nonlinear combinations. PCA is a feature extraction technique, that is used to extract feature subsets from the original datasets to reduce the training time of models in order to achieve dimensional reduction [21]. PCA transforms the original datasets into a subset of principal components (PCs) such that first component of this subset contains the largest amount of information or variance among all components and the last component contains the least amount of information [22].

The aim of this method is to extract the principal components from the original datasets. The pre-processed datasets, $X_{n\times m}$ contains m features and n observation. We construct a $m \times m$ covariance matrix (S), which represents the covariance between each variable. The covariance between two features x_1 and x_2 can be calculated from equation (5)

$$Cov(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2}) \tag{4}$$

$$Covariance\ matrix, S = \begin{pmatrix} Cov(x_1, x_2) & \cdots & Cov(x_1, x_m) \\ \vdots & \ddots & \vdots \\ Cov(x_m, x_1) & \cdots & Cov(x_m, x_m) \end{pmatrix}$$

where, n represents the total number of observations in datasets, $\overline{x_1}$ and $\overline{x_2}$ represents the mean value of features x_1 and x_2 respectively.

The eigen vector and eigen values were calculated to correspond to the covariance matrix (S) and then arranged according to their corresponding eigen values from high to low. We choose k eigen vectors with greatest eigen values and matrix of these eigen vectors is eigenspace represented by $W_{m\times k}$. Furthermore, the principal components are the columns of the datasets (D) which is obtained from the formula (6)

$$\textbf{New dataset}, \mathbf{D^T} = \mathbf{W^T} * \mathbf{X^T} \quad (5)$$

4.3 K-Fold Cross Validation

The cross-validation technique is used to evaluate the performance of ML models by training them on different subsets of available input data and evaluating them on complementary data subsets [23]. In k -fold cross validation, the input datasets is divided into k subsets /folds. Then, we train the ML model on $(k-1)$ folds of datasets and evaluate it on the remaining folds of the datasets by repeating the process k times. Then, an overall performance estimate for the model is obtained by averaging the performance scores from all k iterations. This also reduces the risk of over-fitting.

4.4 Performance Evaluation

Evaluating the performance of regression ML models is a key step in determining their ability to make precise predictions. The most often used metrics for assessing the performance of ML models are R-squared (R^2) mean squared error (MSE).

Mean squared error (MSE) is defined as the average squared difference between the predicted and observed values and is used to measure residual variance [24]. The lower value of this metric represents the good performance of the model. It is determined by using equation (7).

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n} \quad (6)$$

where n denotes the number of data points, $\hat{y}$ denotes the value predicted by a model and y_i denotes the actual value.

R^2 score is a coefficient of determination and is defined as the proportion of the variation in the target variable that can be predicted by the independent variable [25]. It tells how effectively a line fits within a dataset. The value of this score ranges between 0 and 1. This score close to one represents the best fit model for the given problem. This metric is determined by equation (8).

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(y_i - y_{i'})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right) \quad (7)$$

where, y_i represents the actual value, $y_{i'}$ represents the predicted value and $\bar{y}$ represents the mean of target variable.

4.5 Datasets

In this paper, we have used the datasets obtained from Kaggle [40] to predict the patient's LOS in hospitals at the time of admission. The datasets contain one million records in 28 columns with numerical and categorical attributes. This datasets consists of patient medical histories (pneumonia, depression, asthma, psychological disorders, malnutrition, hemo, substance dependence, fibrosis, dialysis, psychother, and iron deficiency), patient demographic information (gender, visit date, discharge date, and readmission rate), vital signs (BMI, respiration, pulse), secondary diagnosis (nonicd9), and lab test results (bloodureanitro, glucose, creatinine, hematocrit, sodium, neutrophils). The total number of days a patient spends in the hospital (length of stay) before being discharged is also included in the datasets, which is a maximum of 17 days. Patients with different medical histories have different LOS and average stay for patients is 4 days shown in Figure 2a.

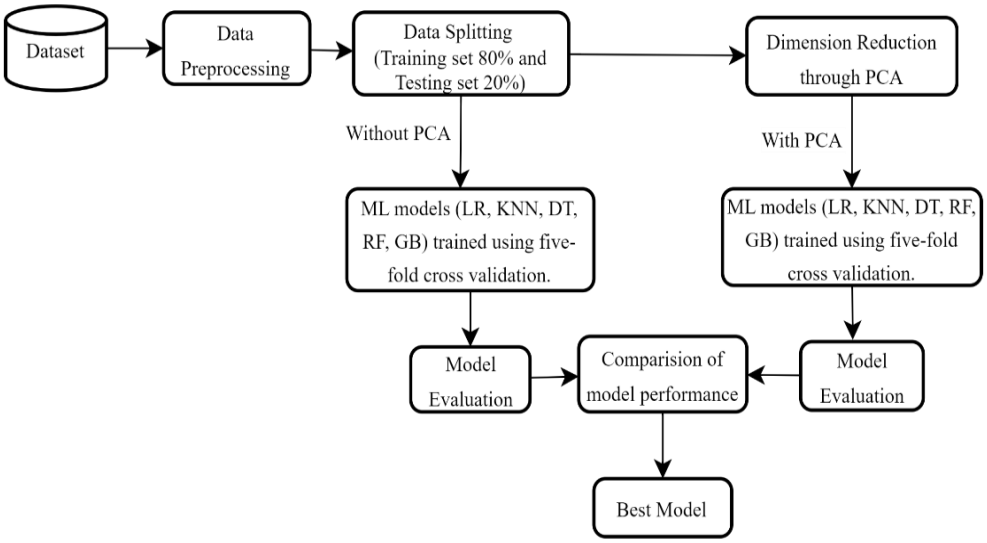


Figure 1. Flow chart of proposed model

4.6 Data Preprocessing

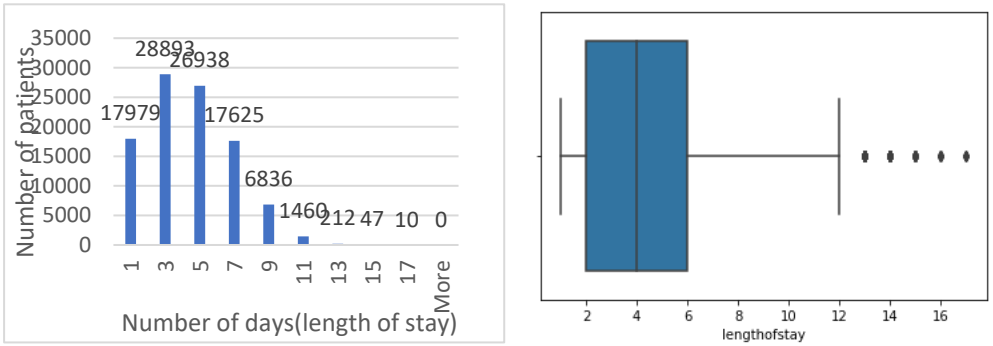
Data preprocessing is needed to enhance the quality of the data and ensure the reliability of the analysis results [41]. In this study, we have applied: (1) outlier removal; (2) data transformation; and (3) data scaling to process the data.

4.6.1 Outlier checking

The datasets are free of missing values. Outliers are the unusual values in our datasets that can distort statistical analysis [42]. If some data is not in range of the main trend, then skewness results, affecting the mean and standard deviation of the distribution. The LOS variable is not normally distributed, so a boxplot is plotted to check for outliers as shown in Figure 2b.

There are some outliers in the right region that need to be removed. The interquartile range (IQR) [43] method is used to remove the outliers.
We calculated the 75th and 25th quartiles using the quartile function and removed the outliers by calculating the upper and lower limits using equations (9), (10), and (11). Therefore, any LOS greater than 12 days is removed from the datasets as an outlier.

$$\text{IQR} = 75\text{th quartile}(Q_3) - 25\text{th quartile}(Q_1) \quad (8)$$
$$\text{Upper limit} = Q_3 + 1.5 * \text{IQR} \quad (9)$$
$$\text{Lower limit} = Q_1 - 1.5 * \text{IQR} \quad (10)$$



a) b)

Figure 2. a) Frequency distribution of LOS with number of patients, and b) Box plot of LOS distribution for outlier detection

4.6.2 Data Transformation

The data transformation technique converts the categorical features to numerical features by using encoding method [41]. Distance based algorithms such as LR and KNN cannot work on categorical data, so one hot encoding technique is used to transform the categorical features such as facility id and gender into a numerical binary column [44]. This method transforms the feature into a binary feature for each category of column, such as gender, which is converted into two columns: male and female.
The dimension of the datasets increased with the encoding method. However, some irrelevant features, such as visiting date, discharge date, and patient id were removed from the datasets. The sample of the datasets after transformation is shown in Figure 3.

facid_A	facid_B	facid_C	facid_D	facid_E	gender_F	gender_M
0	1	0	0	0	1	0
1	0	0	0	0	1	0

Figure 3. Data transformation after applying one hot encoding.

4.6.3 Data Scaling

Data scaling is the third step that must be taken, particularly when the input variables exhibit various scales. Before applying data-scaling, splitting of the data is required to avoid data-leakage. So, datasets are split into a training set (80%) and a test set (20%). The training test is used to train the model, while the test set is held out and not used during the training phase. Scaling is done on both subsets of datasets individually. This step is essential for ensuring the correctness and reliability of the model's predictions. Distance-based algorithms such as LR and KNN are most affected by the range of features. So, the min-max normalization method is used to re-scale the values of all independent features [45]. The re-scaled value ranges between 0 and 1 after data scaling. The formula for normalization is given in equation (11).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{11}$$

where, x_{max} and x_{min} are the maximum and minimum values of the features, respectively.

4.7 Model Development

Subsequent to the preprocessing steps, the pre-processed datasets are utilized for model development. The pre-processed training and testing datasets are used to train and test the model respectively. Furthermore, to evaluate the model's robustness, five-fold cross-validation is applied to the training set. In this study, five ML models were considered: LR, KNN, DT, RF, and GB. Each model was run in two steps to develop the model. First, the models were developed using a pre-processed datasets with 32 features as input. Subsequently, hybrid models were developed using the principal components derived from the pre-processed datasets using PCA.

4.7.1 ML models developed without using PCA approach.

All models LR, KNN, DT, RF, and GB were applied to the training datasets with five-fold cross-validation. We divided the training datasets into five equally sized subsets. Then, each model was trained five times, using four subsets as the training datasets and the remaining subset as the test datasets to assess its performance. Following the completion of the 5-fold cross-validation process, the performance metrics R-squared and MSE obtained from each fold were averaged to produce a more accurate prediction of the model. Then, after training, we evaluated the overall performance of the models on the test set.

4.7.2 ML models developed with PCA approach.

We applied PCA to a pre-processed datasets to reduce its dimension without removing features. We implemented PCA and obtained the principal components (PCs) that represented the original datasets with minimal information loss. The cumulative variance of principal components with increasing numbers of components is shown in Figure 4. Table 1 shows that

first 10 PCs captured 80% variance of all features, 15 PCs contains 90% variance, 20 PCs captured, and 25 PCs components have 100% variance of all features. The contribution of each feature in PCs is shown in appendix A. Afterward, the ML models applied on training datasets with different number of PCs with fivefold cross validation and then evaluated the results on test set.

Table 1. Cumulative explained variance ratio.

```
[ 23.34  40.29  53.07  58.55  63.29  67.47  71.11  74.53  77.83
80.94   83.81  86.    87.83  89.6   91.2   92.67  93.93  95.13
95.98   96.63  97.22  97.76  98.29  98.81  99.29  99.66  99.84
99.97   99.99  100.  100.   100.  ]
```

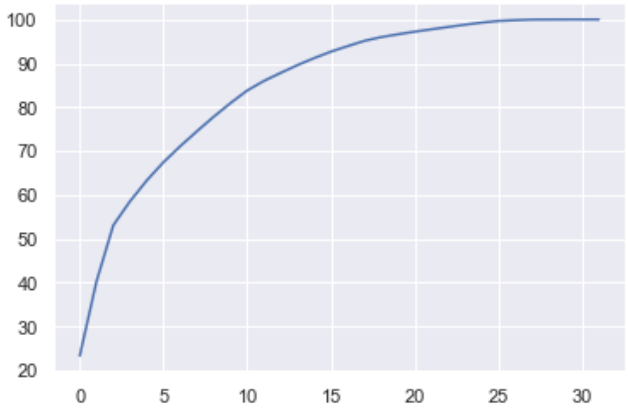


Figure 4. Cumulative variance of components with increasing number of PCs.

5. EXPERIMENT AND ANALYSIS

In this section, the model's evaluation results are presented. Also, the models developed with the original features and those developed with principal components (PCs) have been compared with respect to their overall performance. Regression metrics such as MSE and R- squared values are used to evaluate the performance of all developed models. The results for all developed models on datasets with and without PCA are summarized in Table 2 and 3 respectively. For comparison, R^2 value and MSE value of developed models are shown in Figure 5 and 6.

Table 2. Performance measures of developed models without PCA

Models	MSE	R^2
LR	1.302	0.758
KNN	1.423	0.735
DT	1.428	0.734
RF	**0.775**	**0.856**
GB	0.780	0.855

Table 3. Performance measures of developed models with PCA

Models	20 components		25 components		30 components	
	MSE	R^2	**MSE**	R^2	**MSE**	R^2
LR	1.329	0.753	1.329	0.753	1.302	0.758
KNN	1.499	0.722	1.465	0.728	1.442	0.732
DT	1.893	0.648	1.360	0.747	1.166	0.783
RF	**0.990**	**0.816**	**0.654**	**0.878**	0.523	0.903
GB	1.054	0.804	0.670	0.875	**0.496**	**0.908**

Based on the accuracy results, it can be concluded that the RF model outperformed all the others with greatest R^2 score equal to 0.856, lowest MSE value of 0.775. Furthermore, the results of ML models applied to datasets containing principal components are observed. RF model outperformed all other models with 20 and 25 PCs with R^2 score equal to 0.816 and 0.878 respectively. While GB model performed better with 30 PCs compared to other models. Except for LR and KNN, the accuracy of all models increased as the dataset's dimension decreased. When we used datasets with 20 PCs, then the accuracy results for all models were not good as compared to results of all models without PCA.

Furthermore, the accuracy of all models with 25 PCs and 30 PCs increased as the 99% variation captured by 25 components and 100% variation captured by 30 components. The R^2 score of RF model increased by 2.2%, while MSE decreased by 12% when models developed with 25 PCs. Similarly, R^2 score of RF model increased by 4.7 %, while MSE decreased by 25 % when model developed with 30 PCs. Furthermore, the R^2 score of the LR model on all datasets is almost comparable, such as 0.76 with original features, 0.75 with 10 PCs, 0.75 with 20 PCs and 0.76 with all PCs. The accuracy of KNN model is slightly decreases when integrated with PCA.

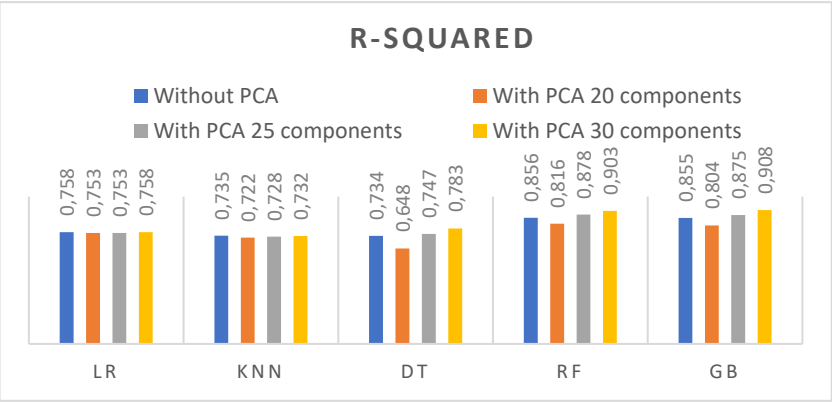


Figure 5. R^2 −score of different ML models with different features.

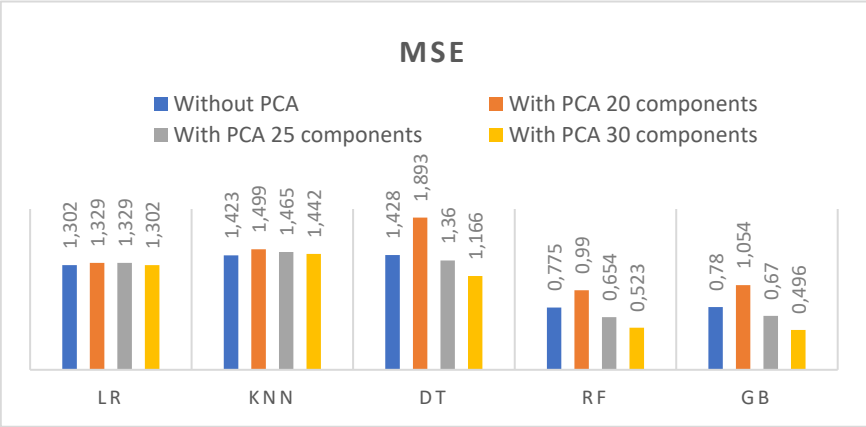


Figure 6. MSE values of different ML models with different features

Furthermore, DT, RF, and GB models performs effectively with PCA, such that it performs better on datasets with PCs greater than 20 components. The accuracy of DT, RF, and GB models increased when these models were integrated with PCA. These models are performing better with PCA as compared to models developed without using PCA. PCA is a transformation technique that performs a linear transformation of the data to a new coordinate system. The KNN is a nonparametric model that makes no assumptions, while LR is a linear model that presupposes a linear relationship between dependent and independent features. As a result, these models evaluated comparable results with and without using PCA. Meanwhile, tree-based models like DT, RF, and GB, capable of capturing both linear and nonlinear relationships, showcase enhanced performance with PCA.

With the respect to all experimental results, it is seen that RF model gives higher performance with the datasets when compared with all classifiers and its performance is improved after using PCA. The GB model performed better with 30 PCs as compared to RF model. But there is very slighter difference between the MSE of both models. The process of dimensional reduction is required for removing the redundancy in data that reduces the time and storage. Therefore, accuracy of all models with PCs lesser than 20 is not so good, while accuracy of all models is increased with 25 PCs. As the cumulative explained variance ratio of PCs increased shown in Figure , R^2 score of tree-based model increased approximate to 0.90 and MSE increased approximate to 0.5.

5.1 Discussion

In this study, our aim was to develop a ML model to predict the LOS of patients and analyse the impact of PCA on the performance of ML models. Therefore, we evaluated the performance of all proposed ML models with and without PCA. LR and KNN models gives comparable result with and without PCA technique, which is satisfactory with previous studies [44], [45]. Cha et al. [46] developed PCA-KNN, PCA-LR, and PCA-DT models using 13 PCs to predict

demolition-waste generation rates in redevelopment areas. They proposed a PCA-KNN model with the highest R^2-score of 0.897. In this study, we found that DT, RF, and GB models performed better with PCA technique, which is consistent with their findings. The accuracy of our proposed models DT, RF, and GB increased when integrated with PCA. Yao [47] developed LR, KNN, SVM, RF, DT, and XGB models with PCA to predict the occurrence of diabetes using health indicators. They observed that XGB model performed best both with and without PCA. However, the accuracy of XGB, SVM, and KNN models decreased when PCA was applied, while the accuracy of LR, DT, and RF models slightly increased with PCA.

In our study, we found that the DT model and RF model performed with better accuracy, when the PCA technique was applied compared to when it was not used. However, the accuracy of KNN model slightly decreased when integrated with PCA. Its accuracy decreases with the loss of information on PCs. Gupta et al. [39] developed a disease prediction model on a Parkinson's datasets using RF and artificial neural network (ANN) models. They also applied the PCA technique to reduce the dimensionality of the datasets and extract the relevant features. The RF model outperformed ANN without PCA, achieving an accuracy of 89%. However, when they applied PCA, the accuracy of RF model decreased by 76%, while the accuracy of ANN model increased from 79% to 97%. This highlights the impact of PCA on different ML models and the potential loss of information.

Additionally, the components obtained through PCA represent a linear combination of original features, which may limit their ability to capture nonlinear relationships among variables in the datasets. Consequently, models incorporating PCA may exhibit sub-optimal performance due to the loss of information from datasets features. In Appendix A, it is illustrated that features like psychother and malnutrition make more significant contributions, whereas gender and facility ID contribute less in PC1, PC2. Consequently, a reduction in accuracy is observed with a smaller number of principal components, while accuracy improves with an increased number of PCs. .All proposed ML models in our study proved their efficiency in previous studies [30], [35], [38]. The proposed DT, RF and GB models are the best prediction model in our study and yielded better results compared to a previous study [48]. Each variable has a different contribution to every principal component. Due to this, not all variables have an equal impact on each component, suggesting that information captured by each PC is not complete. Consequently, the ML models integrated with different number of PCs gives different results.

Our study shows that ML models mainly tree based models when integrated with PCA performed better and evaluate good results. Notably, certain features played a significant role in the combination of PCs. Features such as readmission rate, and medical history (dialysisrenalendstage, asthma, irondef, pneum, substancedependence, psychologicaldisordermajor, depress, psychother, fibrosisandother, malnutrition, hemo, and hematocrit)

demonstrated significant contribution to the PCs. While laboratory tests such as sodium, glucose, bloodureanitro, creatinine, bmi, and PULSE exhibited low contributions to the PCs mainly in first 10 PCs. It is worth mentioning that previous studies [49], [50] have discussed the impact of these lab results on the LOS, further emphasizing their importance. With these findings, limitation in this study is that the models integrated with lesser number of PCs did not lead to improved performance in predicting the LOS of patients during admission. It has been expected that the lower performance was caused by a combination of variables, like a disparity in features and the inadequate information that was acquired by the PCs. In the future there is a need to explore alternative dimensional reduction techniques that account for the varying importance of features and their impact on the prediction task. The study underscores the need to carefully consider the contribution of features to the PCs and their potential impact on the predictive performance of the models.

6. CONCLUSION

In this study, our objective is to develop a ML model for predicting the LOS of patients with severe conditions prior to admission to the hospital. We used the dataset obtained from Kaggle for prediction, and pre-processed it by removing the outliers, transforming the features, and standardising the data. We implemented five ML algorithms, namely LR, KNN, DT, RF, and GB, on a pre-processed dataset with 5-fold cross validation to predict the LOS of patients. We also used PCA as a dimensionality reduction technique to enhance the performance of ML models. All ML algorithms implemented on the datasets with and without PCA. Furthermore, we performed a full evaluation of each model's performance to assess its accuracy in predicting the LOS. The LR model's accuracy is 75%, which is equivalent in both scenarios with and without PCA. The accuracy of the KNN model is slightly decreased with PCA from 0.735 to 0.732, while the accuracy of the DT, RF, and GB models increased when using PCA. However, the RF model performed with the highest R^2 score 0.856 among all models without applying PCA. And models RF and GB integrated PCA performed better with R^2 score approximate to 0.90 and MSE approximate to 0.5. Despite the potential benefits of PCA in reducing the dimensionality and improving model accuracy, our findings also show that applying PCA lead to a significant increase in the predictive performance of the ML algorithms except KNN. The models integrated with PCA with less PCs are not performing well due to the loss of information and the nonlinear nature of the relationships between variables in the dataset. The predictive models developed in our research can serve as valuable decision-support tools for healthcare providers and administrators to estimate the LOS of patients prior to admission, allowing for improved resource allocation and patient management.

It is important to note that the choice of ML algorithms and dimensionality reduction techniques may vary depending on the specific types of datasets and

the nature of the LOS prediction task. In the future, the researchers could explore alternative dimensionality reduction methods and advanced ensemble methods to achieve higher accuracy.

REFERENCES

[1] Oksuzyan A, Höhn A, Pedersen JK, Rau R, Lindahl-Jacobsen R, Christensen K. **Preparing for the future: The changing demographic composition of hospital patients in Denmark between 2013 and 2050**. PLoS One, Vol.15, pp. 1–12, 2020, doi: 10.1371/journal.pone.0238912.

[2] Guidet B, van der Voort PHJ, Csomos A. **Intensive care in 2050: healthcare expenditure**. Intensive Care Med, Vol. 43, pp. 1141–1143, 2017, doi:10.1007/s00134-017-4679-2.

[3] Bsbiology VJC, Cristian A. **Inpatient Rehabilitation Outcome Measures in Persons With Brain and Spinal Cord Cancer**. Cent Nerv Syst Cancer Rehabil 2019.

[4] Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA. **A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients**. Proc - 2014 13th Int Conf Mach Learn Appl ICMLA 2014 2014; pp. 428–431, 2014, doi:10.1109/ICMLA.2014.76.

[5] Mitchell R, Banks C. **Emergency departments and the COVID-19 pandemic: Making the most of limited resources**. Emerg Med J, Vol. 37, pp. 258–259, 2020, doi:10.1136/emermed-2020-209660.

[6] Nhdi N Al, Asmari H Al, Thobaity A Al. **Investigating indicators of waiting time and length of stay in emergency departments**. Open Access Emerg Med Vol. 13, pp. 311–318, 2021, doi:10.2147/OAEM.S316366.

[7] Zhuang Z, Cao P, Zhao S, Han L, He D, Yang L. **The shortage of hospital beds for COVID-19 and non-COVID-19 patients during the lockdown of Wuhan, China**. Ann Transl Med, Vol. 9, pp. 200–200, 2021, doi:10.21037/atm-20-5248.

[8] Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. **Analysis of length of hospital stay using electronic health records: A statistical and data mining approach**. PLoS One, Vol. 13, pp.1–16, 2018, doi: 10.1371/journal.pone.0195901.

[9] Lequertier V, Wang T, Fondrevelle J, Augusto V, Duclos A. **Hospital Length of Stay Prediction Methods: A Systematic Review**. Med Care, Vol. 59, pp. 929–938, 2021, doi:10.1097/ MLR.0000000000001596.

[10] Mittal H, Sharma N. **A Probabilistic Model for the Assessment of Queuing Time of Coronavirus Disease (COVID-19) Patients using Queuing Model**. Int J Adv Res Eng Technol., Vol.11, pp. 22–31, 2020, doi:10.34218/IJARET.11.8.2020.004.

[11] Khosravizadeh O, Vatankhah S, Bastani P, Kalhor R, Alirezaei S, Doosty F. **Factors affecting length of stay in teaching hospitals of a middle-income country**. Electron Physician, Vol. 8, pp. 3042–3047, 2016,

doi:10.19082/3042.
[12] Maulud D, Abdulazeez AM. **A Review on Linear Regression Comprehensive in Machine Learning**. J Appl Sci Technol Trends, Vol.1, pp.140–147, 2020, doi:10.38094/jastt1457.
[13] Uddin S, Haque I, Lu H, Moni MA, Gide E. **Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction**. Sci Rep., Vol. 12, pp.1–11, 2022, doi:10.1038/s41598-022-10358-x.
[14] Nsenge Mpia H, Kasolen MK, Baraka VM, **Inipaivudu Baelani N. Stacking Regression-Based Model for Predicting Patient's Length of Stay in a Semi Urban Hospital.** Int J Res Publ. Rev., Vol. 04, pp. :273–285, 2023, doi:10.55248/gengpi.2023.4212.
[15] Biau∗ G. **Analysis of a Random Forests Model**. J Of Machine Learn Res., Vol.13, pp. 1063–1095, 2012.
[16] Wu Y. **Linear regression in machine learning**. Anal Vidhya, Vol. 161, 2022, doi:10.1117/12.2628053.
[17] Timbers T, Trevor C, Lee M, Peng R. C**hapter 7 Regression I: K-nearest neighbors | Data Science**. Chapter 7 Regres I K-Nearest Neighbors | Data Sci n.d. https://datasciencebook.ca.
[18] Goantiya R. **Tree Based Modeling Techniques Applied to Hospital Length of Stay**. Rochester Inst Technol., Vol. 81, 2018.
[19] Ali J, Khan R, Ahmad N, Maqsood I. **Random forests and decision trees**. IJCSI Int J Comput. Sci Issues Vol. 9, pp. 272–278, 2012.
[20] Aziz N, Akhir EAP, Aziz IA, Jaafar J, Hasan MH, Abas ANC. **A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems**. 2020 Int Conf Comput Intell ICCI 2020 pp.11–16, 2020, doi:10.1109/ICCI51257.2020.9247843.
[21] Zhang C, Cao L, Romagnoli A. **On the feature engineering of building energy data mining**. Sustain Cities Soc., Vol. 39, pp. 508–518, 2018, doi:10.1016/j.scs.2018.02.016.
[22] Sophian A, Tian GY, Taylor D, Rudlin J. **A feature extraction technique based on principal component analysis for pulsed Eddy current NDT**. NDT E Int., Vol. 36, pp. 37–41, 2003, doi:10.1016/S0963-8695(02)00069-5.
[23] Rodríguez JD, Pérez A, Lozano JA. **Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation**. IEEE Trans Pattern Anal Mach Intell, Vol. 32, pp. 569–575, 2010, doi:10.1109/TPAMI.2009.187.
[24] Binieli M. **Machine learning: an introduction to mean squared error and regression lines,** pp. 1–21, 2020.
[25] Chicco D, Warrens MJ, Jurman G. **The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation**. PeerJ Comput Sci., Vol, 7, pp.1–24, 2021, doi:10.7717/PEERJ-CS.623.
[26] Gutierrez JMP, Sicilia MA, Sanchez-Alonso S, Garcia-Barriocanal E. **Predicting Length of Stay across Hospital Departments**. IEEE Access,

Vol.9, pp. 44671–44680, 2021, doi:10.1109/ ACCESS.2021.3066562.
[27] Andersson O. **Predicting Patient Length Of Stay at Time of Admission Using Machine Learning**. Stock SWEDEN 2019.
[28] Gentimis T, Alnaser AJ, Durante A, Cook K, Steele R. **Predicting hospital length of stay using neural networks on MIMIC III data**. Proc - 2017 IEEE 3rd Int Conf Big Data Intell Comput n.d., pp. 1194–1201, 2017, doi:10.1109/DASC-PICom-DataComCyberSciTec.2017.191.
[29] Hijry H, Olawoyin R. **Application of machine learning algorithms for patient length of stay prediction in emergency department during hajj**. Proc Annu Conf Progn Heal Manag Soc PHM 2020, June 2020, doi:10.1109/ICPHM49022.2020.9187055.
[30] Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T, Koblar S. **Machine learning in the prediction of medical inpatient length of stay**. Intern Med J Vol. 2022, pp. 52:176–185, doi:10.1111/imj.14962.
[31] Naemi A, Schmidt T, Mansourvar M, Ebrahimi A, Wiil UK. **Quantifying the impact of addressing data challenges in prediction of length of stay**. BMC Med Inform Decis Mak Vol. 21, pp. 1–13, 2021, doi:10.1186/s12911-021-01660-1.
[32] Siddiqa A, Zilqurnain Naqvi SA, Ahsan M, Ditta A, Alquhayz H, Khan MA, et al. **Robust length of stay prediction model for indoor patients**. Comput Mater Contin., Vol. 70, pp. 5519–5536, 2022, doi:10.32604/cmc.2022.021666.
[33] Aghajani S, Kargari M. **Determining Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department**. Hosp Pract Res., Vol. 1, pp. 51–56, 2016, doi:10.20286/hpr-010251.
[34] López-cheda A, Jácome M, Cao R, Salazar PM De. **Estimating lengths-of-stay of hospitalised COVID-19 patients using a non-parametric model: a case study in Galicia (Spain),** 2021.
[35] Chen Y. **Prediction and Analysis of Length of Stay Based on Nonlinear Weighted XGBoost Algorithm in Hospital**. J Healthc Eng 2021;2021, doi:10.1155/2021/4714898.
[36] MEKHALDI RN, CAULIER P, CHAABANE S, CHRAIBI A, PIECHOWIAK S. **A comparative study of machine learning models for predicting length of stay in hospitals**. J Inf Sci Eng., Vol. 37, pp.1025–1038, 2021, doi:10.6688/JISE.202109_37(5).0003.
[37] Adawiyah R, Badriyah T, Syarif I, Rabiatul Adawiyah, Badriyah T, Syarif I. **Hospital Length of Stay Prediction based on Patient Examination Using General features**. Emit Int J Eng Technol., Vol. 9, pp. 169–181, 2021, doi:10.24003/emitter.v9i1.609.
[38] Wan Z, Xu Y, Šavija B. **On the use of machine learning models for prediction of compressive strength of concrete: Influence of dimensionality reduction on the model performance**. Materials (Basel), Vol.14, pp.1–23, 2021, doi:10.3390/ma14040713.
[39] Gupta I, Sharma V, Kaur S, Singh AK. PCA-RF: **An Efficient Parkinson's**

Disease Prediction Model based on Random Forest Classification 2022.
[40] Choudhury A. **Hospital Length of Stay Dataset Microsoft 2022**. https://www.kaggle.com/datasets/aayushchou/hospital-length-of-stay-dataset-microsoft.
[41] Fan C, Chen M, Wang X, Wang J, Huang B. **A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data**., Front, Vol. 9, pp.1–17, 2021, doi:10.3389/fenrg.2021.652801.
[42] Yusuf AB, Dima RM, Aina SK. **Optimized Breast Cancer Classification using Feature Selection and Outliers Detection**. J Niger Soc Phys Sci., Vol. 3, pp. 298–307, 2021, doi:10.46481/jnsps.2021.331.
[43] Gulati A. **Dealing with Outliers Using the IQR Method - Analytics Vidhya**. Anal Vidhya 2022.
[44] Pei J, Lin X, Chen Q. **Prediction of Patients ' Length of Stay at Hospital During COVID-19 Pandemic Prediction of Patients ' Length of Stay at Hospital During COVID-19 Pandemic**, pp. 0–10, 2021, doi:10.1088/1742-6596/1802/3/032038.
[45] Bhandari A. **Feature Engineering: Scaling, Normalization, and Standardization (Updated 2023).** Anal Vidhya, Vol. 03, Apr 2020.
[46] Cha GW, Choi SH, Hong WH, Park CW. **Developing a Prediction Model of Demolition-Waste Generation-Rate via Principal Component Analysis**. Int J Environ Res Public Health, Vol. 20, 2023, doi:10.3390/ijerph20043159.
[47] Yao L. **Improved Models for Diabetes Prediction by Integrating PCA Technique**, Vol. 47, pp. 106–115, 2023.
[48] Mekhaldi RN, Caulier P, Chaabane S, Chraibi A, Piechowiak S. **Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting**. World Conf Inf Syst Technol., Vol. 1159, pp. 202–211, 2020, doi:10.1007/978-3-030-45688-7_21.
[49] Chuang M Te, Hu YH, Lo CL. **Predicting the prolonged length of stay of general surgery patients: a supervised learning approach**. Int Trans Oper Res., Vol. 25, pp.75–90, 2018, doi:10.1111/itor.12298.
[50] Abd-Elrazek MA, Eltahawi AA, Elaziz MHA, Abd-Elwhab MN, Abd Elaziz MH, Abd-Elwhab MN. **Predicting length of stay in hospitals intensive care unit using general admission features**. Ain Shams Eng J., Vol.12, pp. 3691–3702, 2021, doi:10.1016/j.asej.2021.02.018.

Statement and Declaration:
All the authors/co-authors hereby declare that this research work is original research work, and no financial and non-financial interest is directly or indirectly associated with this work submitted for the publication.

Appendix A

Table A1: Contribution of each feature in each Principal Component.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
rcount	0.001053	0.009378	0.00021	0.02295	0.008696	0.016338	0.081401	0.011635	0.01637	0.000651
dialysisrenalendstage	0.000232	0.004622	0.005718	0.011597	0.005356	0.014629	0.072196	0.017301	0.010617	0.000375
asthma	6.08E-04	2.48E-08	5.74E-05	2.97E-05	1.63E-05	2.44E-05	3.76E-04	2.01E-04	1.78E-05	4.64E-06
irondef	0.000568	0.007274	0.004588	0.034157	0.007865	0.003669	0.017244	0.004312	0.009146	0.000631
pneum	0.000311	0.000021	0.000025	0.000296	0.000299	0.000201	0.000024	0.00018	0.000084	0.000001
substancedependence	4.44E-04	4.02E-04	2.91E-04	1.06E-04	2.78E-05	5.38E-05	1.05E-04	2.43E-04	3.64E-04	3.63E-05
psychologicaldisordermajor	0.002036	0.000352	0.000035	0.000654	0.000267	0.000115	0.000583	0.000375	0.000441	0.000049
depress	0.00418	0.00567	0.00315	0.009622	0.00033	0.01333	0.011055	0.003585	0.011433	0.000085
psychother	0.032691	0.000622	0.00031	0.001444	0.000377	0.002247	0.001803	0.000369	0.001247	0.000031
fibrosisandother	0.001418	0.002106	0.012714	0.009236	0.008972	0.0064	0.006732	0.001061	0.002548	0.000797
malnutrition	0.000283	0.002532	0.004738	0.003613	0.004411	0.015472	0.009685	0.003188	0.000501	0.000584
hemo	1.18E-04	6.24E-03	2.20E-04	2.61E-03	2.28E-03	8.43E-03	1.02E-03	3.10E-03	1.26E-02	1.49E-04
hematocrit	7.99E-05	5.54E-04	4.83E-04	2.25E-03	2.34E-03	1.27E-03	5.07E-03	1.68E-02	1.77E-03	7.61E-06
neutrophils	1.18E-04	2.26E-04	1.19E-04	2.67E-04	2.14E-04	3.28E-04	5.12E-05	1.42E-04	2.74E-04	9.77E-07
sodium	6.04E-05	2.10E-03	3.96E-03	5.96E-03	9.90E-03	3.21E-03	1.86E-03	4.32E-03	9.96E-04	6.00E-05
glucose	4.03E-05	6.01E-04	2.31E-03	5.36E-03	3.11E-03	3.88E-03	2.33E-03	1.45E-04	4.81E-03	3.25E-05
bloodureanitro	0.000013	0.002851	0.006208	0.003808	0.005211	0.002684	0.001703	0.000885	0.000005	0.000012
creatinine	0.000044	0.010338	0.002427	0.000485	0.002668	0.002813	0.002181	0.00034	0.00182	0.00003
bmi	0.00002	0.001876	0.000536	0.000991	0.000585	0.002976	0.003342	0.000768	0.005118	0.000138
PULSE	6.54E-06	8.76E-05	4.16E-03	3.17E-04	2.80E-03	4.69E-04	5.10E-04	1.10E-04	3.35E-04	7.78E-04
respiration	2.45E-06	3.43E-05	1.88E-05	2.83E-05	1.09E-06	2.06E-05	2.34E-05	7.18E-06	2.91E-06	2.56E-05
secondarydiagnosisnonicd9	9.05E-06	3.81E-06	1.35E-05	6.33E-07	1.29E-05	2.64E-05	3.73E-05	2.67E-06	2.20E-05	1.50E-05

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Ar_weekday	0.000006	0.000024	0.000021	0.000008	0.000034	0.000014	0.000002	0.00002	0.000012	0.000021
Ar_day	1.16E-05	1.16E-05	8.16E-05	1.50E-05	2.62E-05	2.44E-05	2.05E-05	1.53E-05	6.95E-07	4.55E-06
Ar_month	1.43E-05	1.96E-05	7.73E-07	2.72E-07	2.94E-06	6.18E-07	1.11E-05	2.05E-05	1.55E-05	1.23E-05
facid_A	0.000002	0.000125	0.000022	0.000002	0.000101	0.000187	0.00001	0.000045	0.00018	0.000242
facid_B	4.52E-07	2.06E-06	1.17E-04	6.79E-06	7.78E-05	2.92E-06	1.35E-06	3.14E-06	2.02E-05	1.81E-03
facid_C	7.95E-09	5.49E-06	8.28E-06	1.42E-05	2.43E-05	1.15E-05	8.47E-06	4.10E-06	8.33E-06	1.21E-04
facid_D	1.05E-08	4.88E-08	7.93E-07	1.96E-07	7.71E-07	1.56E-07	3.86E-07	4.95E-07	2.35E-07	9.89E-06
facid_E	3.88E-08	1.90E-06	3.40E-08	2.14E-07	1.55E-07	2.93E-07	1.45E-08	8.02E-08	2.49E-06	4.12E-08
gender_F	2.69E-48	1.44E-48	1.43E-47	1.90E-47	1.57E-48	3.57E-49	9.69E-48	3.42E-48	2.38E-47	2.51E-47
gender_M	0.00E+00	3.36E-49	9.77E-50	4.54E-49	3.58E-49	2.99E-49	6.17E-51	6.66E-50	6.64E-50	7.32E-49