**ANALYSIS FOR HOSPITAL’S HEALTH CARE DATA**

**LITERATURE SURVEY**

Team id: PNT2022TMID04319

# Prediction of Length of Stay in the Emergency Department for COVID-19 Patients:

 The Kolmogorov–Smirnov test (K-S test) was used to analyze data normality. We performed a one-way analysis of variance (ANOVA) to determine whether there were any statistically significant differences between the median ED LOS and patient race. The normality results of the K-S test will help us identify the correlation method (i.e., Pearson or Spearman correlation) for feature selection. The K-S test yielded p<0.05 , indicating that the data significantly deviated from a normal distribution.

An oversampling technique called the synthetic minority oversampling technique (SMOTE) was employed to address the imbalanced dataset, where observations from the minority class were randomly duplicated. SMOTE generates synthetic samples from the minority class using information available from the given dataset.

It has an accuracy of around 82%.With further validation, the model and results of this study can serve as an effective decision-support tool to improve healthcare delivery/resource planning and help clinicians develop effective interventions to address patient outcomes (e.g., reducing prolonged LOS).

Prediction framework for COVID-19 patient ED length of stay. Data extracted from the hospital database is pre-processed and cleaned for analysis. Models are developed to predict COVID-19 patients ED LOS.

Decision Tree model to build a tree that identifies all possible attribute combinations from the predictive model, and the proportion of COVID-19 patients within the tree experiencing ED LOS less than or greater than 4 hour was calculated.

# Time-to-event modeling for hospital length of stay prediction for COVID-19 patients.

# Hospital length of stay (LOS) is an important indicator of hospital efficiency, quality of patient care, and operational resilience. Numerous researchers have developed regression or [classification models](https://www.sciencedirect.com/topics/computer-science/classification-models) to predict LOS.

# Luo et al. ([Luo, Lian, Feng, Huang, & Zhang, 2017](https://www.sciencedirect.com/science/article/pii/S2666827022000603" \l "b23)) proposed to use logistic regression and random forests to establish a model to predict the LOS of patients with chronic obstructive pulmonary disease. Tanuja et al. ([Tanuja, Acharya, & Shailesh, 2011](https://www.sciencedirect.com/science/article/pii/S2666827022000603" \l "b32)) compared multi-layer perceptron (MLP), naive Bayes, K-NN, and [decision tree models](https://www.sciencedirect.com/topics/computer-science/decision-tree-model) to predict patients’ LOS. Their results showed that MLP and [naive Bayes models](https://www.sciencedirect.com/topics/computer-science/naive-bayes-model) had the best classification of around 85% accuracy, while K-NN performed poorly with only 63.6% accuracy.

# Patient-specific LOS distributions can be learned by using survival models. The features from the text and images could be very useful in improving the accuracy of LOS prediction, which could be an interest of the future research direction. Continuous-time survival methods show good predictive capability compared with discrete-time models. The patients’ vital signs in the form of numeric data

1. **A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients**

Main aim to develop a machine learning algorithm to predict prolonged LOS, defined as >8 days, for patients admitted with COVID-19. Eight days was selected as the threshold of prolonged length of stay based on magnitude of deviation from mean and median length of stays.

Models evaluated include variations on Elastic-net, gradient boosted trees, random forest, support vector machines, logistic regression, a Eureqa classifier, generalized additive models, a Vowpal Wabbit classifier, K-nearest neighbors classifiers, re-sidual neural network, a Rulefit classifier, and ensemble models, which were a combination of other model.

Models were created to predict short LOS on days 1, 2 and 3 of hospitalization. A total of 42 models were trained on data from 80% of patient population, with all models representing variations on the 12 base models. A total of 966 patients were included in this study: 525 of whom had a LOS of ≤8 days, while 441 patients had an LOS of >8 days. Model accuracy improved from 1-3 days of hospitalisation. The development of machine learning algorithms offer a novel approach to tackling the pressing concern of hospital capacity during the ongoing global pandemic.

This work demonstrates that these algorithms are accurate and can be developed for novel disease states for which clinical knowledge is yet unavailable, enhancing clinicians’ ability to make early determinations.

# Hospital patients’ length of stay prediction: A federated learning approach:

The regression analysis performances of the locally trained models and the server-side model aggregating different number of clients have been compared through various parameter metrics. The findings reveal that, the aggregated model’s [predictive performance](https://www.sciencedirect.com/topics/computer-science/predictive-performance) with [federated learning](https://www.sciencedirect.com/topics/computer-science/federated-learning) is less error-prone, and that the model’s performance improves when more clients’ parameters are integrated on the server side.

Ridge regression model is another type of methodology utilized for machine learning regression problems prediction from calculating the coefficients of multiple-regression models for solving co-linearity For Lasso regression, L1 regularization penalty is used instead of L2 regularization penalty that is used in Ridge regression. it is clear that with less number of client the accuracy is low and as the number of client is increased the accuracy of the model also developed. Linear regression, Lasso regression and Ridge regression models are trained with data.

The three types of regression models are all linear analysis-based prediction models, with intercept and coefficient as the parameters used to transfer the learned results to the server. This kind of cloud based predictive model involving decentralized hospital clients without hampering their confidentiality can be a pioneer study for the practical healthcare industry.

1. **The Cycle of Threshold Measurement for COVID-19 Infection and its Implications in Patient’s Length of Hospital Stay**

Data trends were visually and statistically evaluated for normality. We used regression analysis to describe the relationship between categorical variables and between continuous variables with the main outcome. As a result, COVID-19 has rapidly evolved into a pandemic that has put a severe burden on the healthcare services of some countries, notably beds for both hospitalization and intensive care facilities. We used standard qPCR method the estimate the infection and data were analyzed to evaluate the relationship between Ct values with presented variables using stata statistical software. The viral clearance also inversely associated with the Ct values.

Ct value was inversely associated with hospital stay duration (and time to viral clearance), higher the Ct value is indicative of faster time to viral clearance. This could help to manage the infection and allocation of healthcare staff to handle the situation. The strategy has resulted in detecting many asymptomatic patients, in contrast to other countries where the response has been directed solely towards symptomatic patients, and has generated data that is critical for production of accurate prediction models.

Ct value can predict the duration of hospital stay of COVID-19 infected patients, which also suggests that viral clearance will be faster in higher Ct value patients. This data can be implemented in other country to evaluate the patients stay at hospital and also helpful in managing the healthcare staff in successfully handling the COVID-19 infected patients.