Abstract

Unplanned readmissions and long stays are becoming a major concern of hospitals and healthcare providers as an indicator for quality of service. Predicting readmissions and long hospital stays in early stages allows providing extensive attention to patients identified with higher risks, which leverages the community's healthcare and saves healthcare expenses and resources. The limited ability of conventional statistical techniques presents a challenge to the developing accurate predictive models for readmissions and length of stay. Understanding the impact of each predictor in modelling patient risk factors permits the provision of effective solutions directed to solving crowding issues. This project aims to investigate the predictive factors and develop a robust risk prediction framework, by combining feature engineering and machine learning algorithms. For the model building process, advanced machine learning models, such as support vector machine, random forest, and extreme gradient boosting, are trained with preprocessed medical data. The proposed methodology is applied to an actual dataset with various levels of routinely collected data that includes demographics, admission information, diagnosis, medications, tests, service utilization information. Results obtained from comparative experiments demonstrate the effectiveness of the proposed framework in prediction using a set of elite features, which promotes the generalizability of the framework. A readmission prediction accuracy of 94.8% has been achieved using random forests. The support vector machine (SVM) gives the highest area under curve (AUC) statistic (i.e. 0.97) in readmission prediction, and it stands out as an efficient algorithm in predicting length of stay with prediction accuracy of 78.5%. Furthermore, insightful implications are obtained from the analysis of features and can be used to determine at-risk patients and target the delivery of early resource-intensive interventions.