Journal of Health Informatics and Management

Research Article A SCITECHNOL JOURNAL

An Efficient Model for Predicting High-Cost Members

Yan Xu*, Kirit Pandit, Jason Cooper

Health Management Systems Inc, Irving, Texas, USA

***Corresponding author: Yan Xu,** Health Management Systems Inc, Texas, 75038, USA, Tel: 469-284-6039; E-mail: yan.xu@hms.com

Received date: 6 January, 2021; Accepted date: 23 January, 2021; Published date: 09 February, 2021

Abstract

Healthcare costs have been a significant issue in the United States. A large proportion of the cost is potentially avoidable if high-quality care is provided to the appropriate high-cost members. Moreover, only a small proportion of members consume the majority of healthcare costs and medical resources. By predicting more accurately those likely to become high-cost members and providing proactive care to those members, a large amount of the medical cost and resources can be diverted with effective prevention and outreach programs. Accurately identifying high-cost members in the future is critical for proactive care.

With the rapid advancements of big data platforms, as well as data science tools and techniques, machine-learning algorithms have been successfully applied for predicting different types of outcomes within the healthcare community. In this paper, we propose an efficient and effective predictive model based on the recent open-source algorithm LightGBM to predict various high-cost members based on different types of features. The predictive model was compared to several current "state-of-the-art" and well-known machine-learning algorithms. Additionally, different feature categories were extracted from the claims data, social determinants of health data, access to care information, and the proposed algorithm was evaluated on different feature groups. The experimental results show that the proposed predictive model combined with a specific feature group with more robust features can yield better performance in terms of different metrics while retaining operational efficiency.

Keywords: Healthcare; High-cost; Predictive model; LightGBM

Introduction

In 2018, the healthcare cost in the US consists 17.7% of its gross domestic product, and it is expected to rise to 19.7% in 2028 [1]. On the one hand, there has been an increasing interest to shift from the traditional fee-for-service (providers are paid for the volume of services performed) to the value-based care (providers are paid based on patient health outcomes – value derived from services), which can both provide better care for individuals and lower the cost [2]. The fee-for-service system will result in 'quantity care' while the value-based system is for 'quality care'. It is estimated that the top 20% of the population accounts for 80% of the total expenditures [3]. For those high-cost members, if their costs can be managed well by providing proactive care and outreach programs, then the total expenditures can be reduced. However, predicting the high-cost members is a challenging problem since a member's probability of becoming a high-cost member or a low-cost member is associated with time and the quality of care provided to them. Many factors such as their age, chronic conditions, historical outcomes and gaps in care, etc., will determine if they will become a high- or low-cost member. Moreover, there are several reasons for high-cost, including admissions and readmissions, emergency (ER) visits, medication costs, post-acute care costs like hospice care, home health and others. Therefore, accurately identifying those high-cost members and their future cost using the rich feature information for proactive care is of great importance [4-5].

Artificial Intelligence has gained increasing popularity in the area of healthcare, especially with the evolution of big data platforms. This is because the large volume of data can provide more useful patterns for machine-learning algorithms to adapt to those patterns and separate different classes. Moreover, the electronic health care claims submission process makes it much easier to fully extract the available features and exploit the power of machine learning. Different data sources including medical claims, pharmacy claims, eligibility, and even Electronic Medical Records (EMR) [6] have been used for predicting different types of outcomes. For instance, in [7], a Logistic Regression predictive model is adopted for predicting the binary outcome to enhance the clinical decision-making. Decision tree based algorithms have gained popularity because of their predictive capability. In [8], a decision tree based algorithm is presented to predict the utilization of healthcare based on socioeconomic data. Gradient Boosting Decision Tree algorithms such as XGBoost [9] have been proposed to predict the hospital admission, and the results show it offers superior performance compared with the logistic regression and deep neural network predictive models. Improved from XGBoost, LightGBM [10] proposed novel techniques to reduce the training time and increase the performance.

In this paper, we propose to use LightGBM to predict high-cost members using different set of features that are available on the claims data, Social Determinants of Health (SDoH) and Access to Care (ATC) information. Different metrics including Precision, Recall, F1 score and Area under Curve (AUC) score of the Receiver Operating Curve are reported. To investigate the importance of the different feature categories, the model is tested on different set of features. The organization of the paper is as follows. First, the dataset and the features are discussed. Then, the proposed algorithm and the metrics for performance evaluation is presented. The experimental results are discussed followed by the discussion and conclusion section.

Dataset and Features

The dataset that is used for experimentation in this study is a population in one state of the United States. The dataset that we used is from January 2015 to December 2019. The number of eligible membership for the latest month (December 2019) was 119,168. The average age for this population is 30, and the number of male and female is 55987 and 63182 respectively. The number of members who has enrolled more than six months at December 2019 is 113227, and the number of members who has enrolled more than one year is 91324 Different features are constructed from one-year data, and then the target variable is formed using the costs from the following year. The detailed formulation of the features and targets is shown in Table 1.

After eliminating the members who were enrolled for less than six months in the year of the features, the resulting total number of samples is 319,328 spanning our dataset's four-year timeframe. Eighty percent of the total samples in this dataset are randomly selected for training and cross validation, and the rest of the samples are used for testing. All the performance on different metrics is reported on the testing data. The detailed statistics of the training and the testing dataset are shown in Table 2. Based on the statistics in Table 2, we decided to use the yearly cost above the 80th percentile of the total cost in the training dataset to represent the high cost cohort. The target variable is a binary variable, and the yearly cost above the 80th percentile of total cost in the training dataset is set as one to represent the high-cost members. Otherwise, it is denoted as zero to represent the lower cost cohort. It can also be noted from Table 2 that the distributions between the training and testing data are close.

Features	Targets
2015/01 to 2015/12	2016/01 to 2016/12
2016/01 to 2016/12	2017/01 to 2017/12
2017/01 to 2017/12	2018/01 to 2018/12
2018/01 to 2018/12	2019/01 to 2019/12

Table 1: Features and targets formulation.

Characteristic	Training set	Testing set
Patients, N (%)	255364 (80)	63964 (20)
Age, mean (SD)	32 (21)	32 (21)
Male sex, N (%)	113233 (44.34)	28482 (44.52)
More than five chronic conditions, N (%)	61105 (23.93)	15015 (23.47)
Three to five chronic conditions, N (%)	57476 (22.51)	14461(22.61)
Two chronic conditions, N (%)	32458 (12.71)	8210 (12.84)
One chronic condition, N (%)	44426 (17.40)	11234 (17.56)
No chronic conditions, N (%)	59899 (23.46)	15044 (23.52)
More than five open gaps in care, N (%)	59671 (23.37)	14762 (23.08)
Three to five open gaps in care, N (%)	140057 (54.85)	34761 (54.34)
Two open gaps in care, N (%)	41744 (16.35)	10874 (17.00)
One open gap in care, N (%)	12516 (4.90)	3209 (5.02)
Zero open gaps in care, N (%)	1376 (0.54)	358 (0.56)
Annual cost 80 percentile ($)	6005.3	5986.9
Annual cost 60 percentile ($)	2043.3	2055.8
Annual cost 40 percentile ($)	892.1	900.4
Annual cost 20 percentile ($)	291.3	296.2
More than two hospital admissions, N (%)	177 (0.07)	39 (0.06)
Two hospital admissions, N (%)	1751 (0.69)	422 (0.66)
One hospital admission, N (%)	13739 (5.38)	3428 (5.36)
Zero hospital admission, N (%)	239355 (93.73)	60000 (93.80)

Table 2: Detailed statistics of the training and testing dataset.

To investigate the effects of the features on the model, we separate the features into five different sets. Each set includes different feature categories as shown in Table 3. The total number of features in each set is also shown in Table 3. The feature category includes demographic, SDoH and ATC risk scores, chronic conditions, historical outcomes and indexes.

Feature group	Features category	No. of Features
Set1	Demographic Only	2
Set2	Demographic + SDoH + ATC	4
Set3	Demographic + SDoH + ATC+ Chronic conditions	46
Set4	Demographic +SDoH + ATC + Chronic conditions + Historical outcomes	379
Set5	Demographic + SDoH + ATC + Chronic conditions + Historical outcomes + Indexes	401

Table 3: Different feature groups.

The demographic features include age and gender. We also include the nonclinical feature categories including SDoH risk score and ATC risk score. The SDoH risk score is calculated by using a logistic regression machine learning model, and the features include education level, occupational and financial well-being, physical and social well-being, housing and nutrition, environmental and access-to- care at a census tract level. The ATC risk score is calculated based on the distance between the members' home location and their primary care physician and available hospitals. The chronic conditions category includes forty chronic conditions, two indicators of rare disease and one-time disease. In this category, there are a total of 42 features. For example, if a patient has hypertension, then the corresponding value of this feature will be assigned as one. On the contrary, if a patient does not have hypertension, then the value of the hypertension feature column will be zero. This could also be considered a comorbidity set.

In the historical outcomes category, monthly cost of different types and historical outcomes including number of emergency visits, hospital admissions, office visits, etc., are constructed as a time series.

For instance, monthly inpatient cost for each month is included in the feature matrix for the twelve months. Simple moving average and exponential moving average [11] for different outcomes and cost is calculated for a twelve-month window. The exponential moving average of the monthly cost is a feature that can determine the trends of cost for that member. High value of this feature indicates that this member has higher medical cost or number of outcomes in recent months compared with prior months. The gaps in care for the chronic conditions are also included in those features. For instance, months since the patient was seen for Diabetes, though not a National Committee for Quality Assurance (NCQA) gap, has been considered as a diabetes care persistence gap for purposes of this model, and is considered as one feature. There are a total of 333 features in this category.

In the indices category, LACE index [12], Charlson index [13] Chronic Illness and Disability Payment System (CDPS) risk score [14] are included. Additionally, each ICD diagnosis code and NDC drug code has been assigned a score from the training dataset by grouping the different types of outcomes. Since there are a large number of ICD and NDC codes, this can be considered as a dimensional reduction approach while taking full advantage of the available information. Based on the individual ICD and NDC scores for different types of outcomes, the ICD and NDC index are calculated and formed as features. For instance, if a member has ten ICD codes, then the ICD index for hospital admission is calculated as the summation of the individual ICD scores weighted with respect to hospital admissions associated with that code. Since the outcomes can be different types including hospital admission, Emergency Department visits, Readmission, and the target can be grouped in a period of a year or the last quarter, different ICD and NDC indices can be generated as features. In this category, there are a total of 22 features.

Based on the different category of features, we constructed different set of features as shown in Table 3. For instance, feature set1 only include the demographic features. Feature set2 include the demographic features, SDoH risk score, and the ATC risk score. Feature set3 includes both the demographic features SDoH risk score, the ATC risk score, and the chronic conditions. The Chronic conditions are list in Table 4. In feature set4, demographic, SDoH risk score, ATC risk score, the chronic conditions, and the historical outcome are included. Finally, in the last set, the index features are included along with all the features in set4.

CAD	Hypertension	Diabetes	CHF
Asthma	Joint disorders	Cancer	Cerebrovascular disease
Anxiety disorders	COPD	Prediabetes	Osteoporosis
Depression	Renal disease	Obesity	Arthritis
Epilepsy	Autism	HIV	Substance use disorder
Sickle Cell Anemia	Cerebral Palsy	Congenital anomalies	End stage renal disease
Cystic Fibrosis	Neoplasm	Bipolar disorder	Neuro cognitive disorder
Schizophrenia	Personality disorders	OCD	Neuro development disorder
Hemophilia	Hyperlipidemia	Intellectual disabilities	Organ transplant
Immunity disorder	Seizure disorder	Somatic disorders	Trauma and stress

Table 4: Different chronic conditions included in the features.

Methods and Metrics

The algorithm that we used in the experimentation is LightGBM, which is an Efficient Gradient Boosting Decision Tree. Improved from XGBoost, two techniques including Gradient-based One-Side Sampling and Exclusive Feature Bundling are implemented for high accuracy and scalability [10].

The performance metrics in the experimentation include Accuracy, Area under curve (AUC), Precision, Recall, F1 score. The Accuracy is the total correctly predicted testing samples out of all the testing samples regardless of their classes. AUC is the area under the Receiver operating characteristic curve (ROC), which plots the trade-off of specificity against the sensitivity for every cut-off. The larger this value, the better the performance. For Precision, Recall, F1, we illustrate the cases for the positive class. The rule is the same for the negative class if the class labeled is reversed. Thus, for simplicity, for those three measures, we illustrate the case for the positive class only. Precision is defined as the ratio of the correctly predicted positive samples over all the true positive samples as shown in Equation (1). Recall is defined as the ratio of the correctly predicted positive samples over all the true positive samples as shown in Equation (2). Note Precision for the positive and negative class is denoted as Precision (positive) and Precision (negative) respectively. Recall for the positive and negative class are defined as Recall (positive) and Recall (negative) respectively. Recall (positive) and recall (negative) are also known as Sensitivity and Specificity respectively. The F1 score is a trade-off metric that balances the precision and recall as shown in Equation (3).

Precision=(True Positive)/(True Positive+False Positive) (1)

Recall=(True Positive)/(True Positive+False negative) (2)

F1=(2*Precision*Recall)/(Precision+Recall) (3)

Results

First, we compare the performance of different algorithms using all the available features. The algorithms include Logistic Regression (LR) [15], Naïve Bayes classifier (NB) [16], K-nearest Neighbors algorithm (KNN) [17], Neural Network (NN) [18], Decision Tree algorithm (DT) [19], XGBoost [20], and LightGBM. LR is a linear classifier that aims to maximize the binary cross entropy among the training samples to separate the classes. NB is a simple probabilistic classifier that relies on the Bayes theory. KNN is a non-parametric classifier that in the prediction process, the testing sample's label is determined by its n closest distance from the training samples. NN is the classifier that consists of multilayers of neurons, and during the training process, the weights of the neurons are generated by the training samples. In the prediction process, the probability of belonging to different classes is calculated by the weights and each testing sample feature for each testing sample. A DT classifier generates the tree-like rules during the training process, and the prediction process can be determined by the learned rules from the

training process. XGBoost is a scalable tree boosting algorithm that is easily scalable and has been widely used in the industry. Five- fold cross validation is implemented on all the algorithms for hyper-parameters tuning to select the best parameters for fair comparison. The detailed parameters by cross validation for each algorithm are shown in Table 5, Figure 1 shows the AUC for all the algorithms. We can see that the proposed LightGBM algorithm can yield the best performance in terms of the AUC value. Another observation is that Decision tree based algorithms including LightGBM, XGBoost, DT have better performance in terms of AUC value compared with other algorithms investigated. The NN classifier also has good performance, followed by KNN and Naïve Bayes algorithm. KNN has the worst performance in terms of AUC.

Algorithms	Parameters: value
LR	["C":0.001, "penalty":L2]
NB	["smoothing parameter":1e-8]
KNN	["number of neightbors":5] Journal of health informatics and management
NN	["hidden layer sizes":(100, 2), "activation":'relu']
DT	["criterion":'gini', "max depth": 6]
XGBoost	["learning rate":0.02, "max depth":8, "gamma":0.7, "min child weight":2, "estimators":500, "max leaves":300, "col-sample by tree":0.8]
LightGBM	["learning rate":0.01, "max depth":12, "max le aves":260, "min child weight":3, "estimators":900, "col-sample by tree":0.61, "regularization lambda":0.2]

Table 5: Parameters by cross-validation for different algorithms.

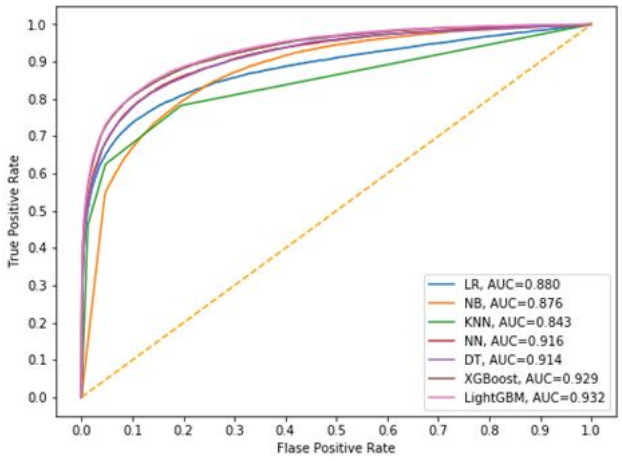

Figure 1: ROC curve for different algorithms.

Table 6 shows the performance of all the algorithms for different metrics. It can be concluded that LightGBM can outperform other State-of-the-Art algorithms including LR, NB, KNN, NN, DT, and XGBoost in terms of Accuracy and AUC. For the metrics Precision (positive) and Recall (negative), KNN has slightly better performance than LightGBM. The reason might be that for KNN, there are more negative samples that have been predicted in the prediction process. Since KNN algorithm determines the class label by calculating the similarity between the testing sample and each training samples, and the most samples in the training dataset belongs to class zero. Therefore, when calculating the similarity in the prediction process, more zero class samples are determined. However, in terms of Recall (positive), which is a more important metric to predict the high-cost members out of the total high-cost members, LightGBM can yield significantly better performance than KNN.

Algorithms	LR	NB	KNN	NN	DT	XGBoost	LightGBM
Accuracy	0.8809	0.8591	0.8876	0.8914	0.8937	0.9014	0.903
AUC	0.88	0.8756	0.8432	0.9159	0.9143	0.929	0.9321
Precision (negative)	0.9261	0.9124	0.9108	0.9327	0.93	0.9389	0.9397
Precision (positive)	0.7004	0.6458	0.7676	0.7266	0.7415	0.7514	0.7545
Recall (negative)	0.9251	0.9114	0.9528	0.9316	0.9378	0.9378	0.9386
Recall (positive)	0.7034	0.6486	0.6251	0.7298	0.7164	0.7547	0.7578
F1(negative)	0.9256	0.911 97	0.9314	0.9322	0.9339	0.9383	0.9391
F1 (positive)	0.7019	0.6472	0.6891	0.7282	0.7287	0.7531	0.7561

Table 6: Different metrics on the algorithms.

Additionally, we compare the performance of the LightGBM using different set of features in Figure 2. The motivation is to investigate if adding new types of features will increase the performance of the model. It can be clearly seen that with only demographic features, the performance is very poor. For example, the Precision and Recall of the positive class is even less than 0.5. With the SDoH risk score and ATC risk score incorporated, the performance of the model can be slightly improved with the demographic only features. By adding the different chronic conditions feature, the performance of the model is increased. If more features including the historical outcomes features are added, then we see a clearly improved performance for all the metrics. The reason is that when the number of features is very low, samples between different classes are very hard to be distinguished with each other in a low dimensional space. By adding more features and increasing the nonlinearity, samples are easier to be separated from each other. Moreover, we could see a slight increase in performance over different metrics when adding the index set of features.

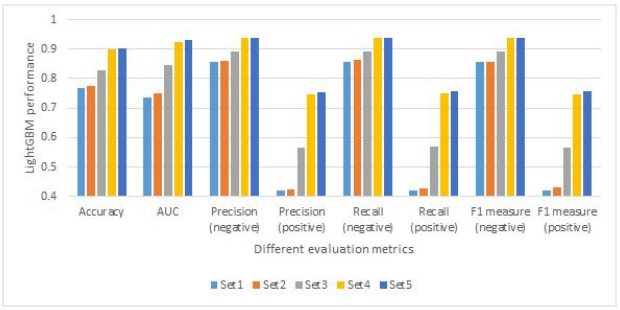

Figure 2: LightGBM performance for different feature groups.

The training and testing time of the different algorithms are presented in Table 7. As we can conclude from this table, the training time for LR, DT, and NB is less compared with XGBoost, LightGBM, and NN. This is expected because LR, DT, and NB are simpler classifiers and have less parameter during the training process. For XGBoost and LightGBM, they are improved from the DT algorithm by using boosting algorithms. Additionally, we can see that XGBoost has a longer training time compared with LightGBM, which is expected. KNN has the longest testing time, and this is because in the testing process, the testing sample needs to compare with every training sample to find the closet neighbors.

Algorithms	LR	NB	KNN	NN	DT	XGBoost	LightGBM
Training time (s)	19.7814	3.15367	-	242.375	15.3721	298.9	172.34
Testing time (s)	0.1785	0.86967	1464.93	0.875629	0.22639	0.001	3.70014

Table 7: Computational cost of different algorithms.

Discussion

The performance of the proposed LightGBM with a set of rich features and the model using the CDPS risk score is also reported in Figure 3. As we can conclude from this figure, the proposed model will significantly outperform the model that applies CDPS risk score only. For instance, for the recall on the positive class, the proposed algorithm is 0.76 while the one using the CDPS risk score is only 0.47, which means that the proposed algorithm can accurately identify more than 29% of the total high-cost members. The AUC score of the proposed model is also significantly larger than the one using the CDPS risk score. Additionally, we investigate the amount of the cost on the testing data using both the proposed model and CDPS risk scores. The total cost of the members on the testing data is $408.16 million. By reaching out to the members with the positive predicted categorization, $311.046 million can be accurately identified by the proposed model, compared with only $127.16 million using the CDPS risk score.

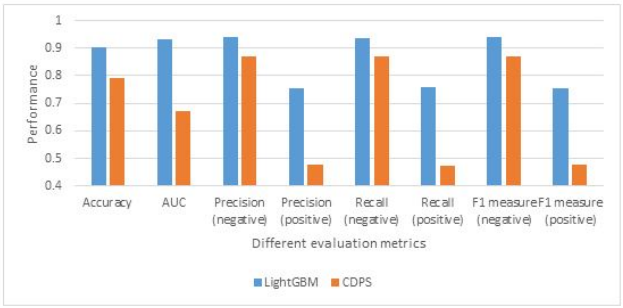

Figure 3: LightGBM and CDPS risk score performance.

Conclusion

In this paper, we consider the use of an efficient algorithm LightGBM for predicting the high-cost members. The proposed algorithm is compared with current state-of-the-art machine learning algorithms over several metrics. The experimental results show that the proposed algorithm consistently outperforms current state-of-the-art algorithms over different metrics. Different sets of features are also investigated on the performance of the proposed model to evaluate the different aspect of features. Additionally, the performance of the proposed machine learning model is compared with the model using CDPS risk scores. From our study and methods utilized, one could conclude that by including more features, the predictive modeling performance can be significantly improved and accurately predict high-cost members.

References

1. Hartman M, Martin AB, Benson J, Catlin A (2020) National health care spending in 2018: growth driven by accelerations in Medicare and private insurance spending. Health Aff 39.
2. NEJM CATALYST, "What is value-based healthcare?" N.Engl.J.Med., 2017
3. Stanton MW, Rutherford M (2006) The high concentration of U.S. health care expenditures. AHRQ Research in Action 19.
4. Yang C, Delcher C, Shenkman E, S Ranka (2018) Machine learning approaches for predicting high cost high need patient expenditures in health care. BioMed Eng OnLine 17.
5. Duncan I, Loginov M, Ludkovski M (2016) Testing alternative regression frameworks for predictive modeling of health care costs. N Am Actuar J 20.
6. Laszczyńska O, Severo M, Azevedo A (2016) Electronic medical record-based predictive model for acute kidney injury in an acute care hospital. Stud Health Technol Inform.
7. Shipe ME, Deppen SA, Farjah F, Grogan EL (2019) Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis.
8. Chen S, Bergman D, Miller K, Kavanagh A, Frownfelter J, Showalter J (2020) Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care. Am J Manag Care.
9. Hong WS, Haimovich AD, Taylor RA (2018) Predicting hospital admission at emergency department triage using machine learning. LOS ONE 13.

10. Ke G, Meng Q, Finley T, Wang T, Chen W, et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inform Process Syst.
11. Klinker F (2010) Exponential moving average versus moving exponential average. Math Semesterber 58.
12. Low LL, Lee KH, Hock Ong ME, Wang S, Tan Sy, et al. (2015) Predicting 30-Day Readmissions: Performance of the LACE index compared with a regression model among general medicine patients in Singapore. Biomed Res Int.
13. Quan H, Li B, Couris CM, Fushimi K, Graham P, et al. (2011) Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstructs using data from 6 countries. AM J Epidemiol 173.
14. Juhnke C, Bethge S, Mühlbacher AC (2016) A review on methods of risk adjustment and their use in integrated healthcare systems. Int J Integr Care 16.
15. Cramer JS (2001) The origins of Logistic Regression. Tinbergen Institute Discussion Papers.
16. Kumar S, Jain A, Mahalakshmi P (2018) Enhancement of healthcare using naïve bayes algorithm and intelligent data mining of social media. Int J Appl Eng Res 13.
17. Sun J, Du W, Shi N (2018) A survey of kNN algorithm. Inf Eng Appl Comput.
18. Shahid N, Rappon T, Berta W (2019) Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLoS One 14.
19. Tanner L, Schreiber M, Low JGH, Ong A, Thomas T, et al. (2008) Decision Tree algorithms predict the diagnosis and outcome of Dengue fever in the early phase of illness. PLos Negl Trop Dis 2.
20. Chen T, Guestrin C (2016) XGBoost A scalable tree boosting system. In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM.