International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Predicting Home Health Care Services Using A Novel Feature Selection Method

C. Arulananthan[1], K. Sujith[2]

[1]Research Scholar, Department of Computer Science, Annai College of Arts & Science (Affiliated to Bharathidasan University, Tiruchirappalli), Kumbakonam, Tamilnadu, India.

[2]Research Advisor & Associate Professor, Department of Computer Science, Annai College of Arts & Science (Affiliated to Bharathidasan University, Tiruchirappalli), Kumbakonam, Tamilnadu, India.

Abstract: In the hospital management and medical areas, the experience of the patients is seen as having a dominant reputation. Online patient reviews are acknowledged as a crucial factor in assessing the effectiveness and quality of healthcare services. Many people find the traditional method of assessing service excellence to be cumbersome. However, the data assessment and evaluation processes are now more informal and time-efficient thanks to machine learning classifiers and opinion mining tools. Patient satisfaction and hospital patient service quality currently play a significant role in the health care industry. This is achieved by foreseeing the various hidden patterns and pinpointing the essential elements that contribute to patient fulfilment. This work provides the novel feature selection method for identifying the key feature in Home Health Care Services using patient satisfaction data. Several components are examined to evaluate the superiority of different health care services based on diverse metrics. Various machine learning algorithms are applied to guess the significant aspects affecting patient healthcare satisfaction. As recognized, the patient experience is an indispensable for assessing the quality in healthcare services.

Introduction

The quality of care offered in any healthcare sector is of paramount importance to all investors in that domain. Different elements are employed to spot the level of consumer satisfaction in the healthcare sector [1]. In all government bodies, health-related well- being is regarded as one of the most vital factors [2]. The various aspects of determining the service quality include effectiveness, safety, communication, and overall care, all of which play a significant role in assessment. Opinion mining is an essential role in judging healthcare quality. It seeks out patient hidden patterns for estimating the consumer satisfaction. In this case, the consumer is the ultimate authority in determining the quality while receiving healthcare services.

Recently, the consumer's opinion on each aspect allied with healthcare has been evaluated through patient reviews on web forums or response forms provided by the healthcare unit. It enables enterprises to advance their services and broaden their consumer-oriented base. The best source of knowledge in evaluating healthcare quality is imported from patient reviews which provides vital data for this research on their experiences.

Previous studies reveal that the higher percentage of patient satisfaction directly impacts the healthcare sector's growth [3], [4]. Moreover, past works observed that identifying the key element in patient satisfaction is very cumbersome [5], [6]. Yet, many researches are venturing into analyzing the factors associated with opinion mining towards patient satisfaction. So, identifying the best feature measure for patient satisfactionis still debatable. So many measurable features can be predicted in varied dimension that take the studies to give many multidimensional conclusions firmly. Studies reveal that onlylimited tools and models are available to measure this satisfaction. The most commonly used assessing tools are survey patients' opinions based on their services. Potential researchis desirable to identify the critical factor in assessing the measurable feature in satisfying the patient that helps the provider to excel in their service [7], [8].

With the traditional tools, it is challenging to identify the varied dimensions of opinion mining that has different associations with varied metrics in assessing patient satisfaction. 'Machine learning classification methods such as LR, RF, GB, and AB are employed to identify the potential features, which aids in predicting the favorable part to evaluate patientsatisfaction. Very fundamentals to assess the patient satisfaction and the services renderedby healthcare lead to the factor of identifying another metric of determining the quality of care in relation to patient opinions. In U.S, the HCAHPS is a review for health care that measures experience of patient with various metrics. It provides a good hospital analysis outcome and different insights for improving service quality [9]. The survey conducted by HCAHPS in various cities across Australia is used to analyze the various home-health

–care-services offered under different vendor bases. This research proposes ML based approach for analyzing patient ranking datasets.

Feature importance helps in identifying the critical feature which is vital for patientsatisfaction [11] [12] [13] [14] [15] [16]. For each model, a classification report is generated. ROC curves is for validatemodel accuracy, and the model cross-validation is evaluated.

Furthermore, this paper is categorized as section 2 which explains the proposed methodology and data collection, section 3 discusses the findings and discussions from the proposed approach, section 4 discusses model validation and performance analysis, and finally section 5 concludes the paper.

2. Proposed Methodology

The framework is widely accepted that the feature selection approach is meant for its effectiveness in choosing the best features and forecasting the model based on the features selected. The proposed work provides a robust classification model for home health care data by guessing the finest result based on the selected features. It also advocates a novel methodology for classification using machine learning models that include a feature selection model and a feature importance model. Fig. 1 depicts the proposed framework. After reading the data with Python libraries, the data is preprocessed by deleting missing values and formatting the data with the filter method.

The feature importance is used to achieve robust prediction based on the model assessed. Subsequently, machine learning (ML) classifier is employed to choose the premium approach after obtaining unique features.

2.1. Implementation

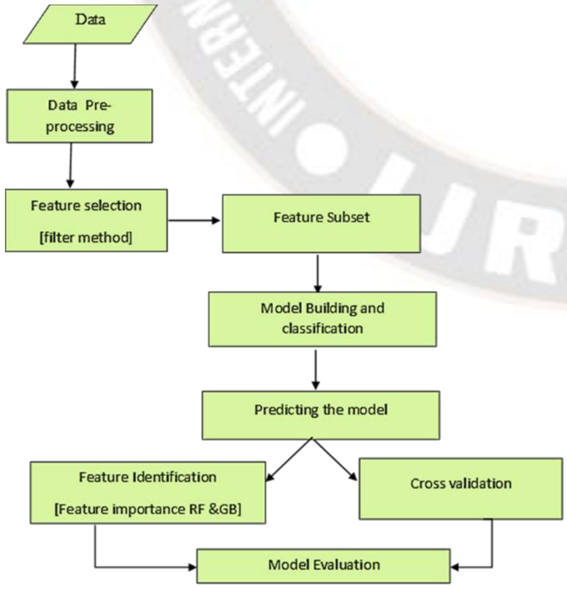
Data
Data Pre-processing
Feature selection [filter method]
Feature Subset
Model Building and classification
Predicting the model
Feature Identification [Feature importance RF &GB]
Cross validation
Model Evaluation

Fig. 1.: Proposed Framework

Implementation of the proposed framework is explained below for home health care prediction using a filter-based feature selection process, along with the proposed procedure for implementing the methodology mentioned above.

The sequence of procedural steps

1) *Begin*
2) *Import data for home healthcare.*
3) *Implement Data Pre-Processing.*
4) *Create the Feature Subset FS.*
5) *Gather test and training data.*
6) *Build the model RF, GB and AdaBoost*
7) *Generate Feature Importance using RF, GB and AdaBoost*
8) *Using the ML classifiers RF, GB, and AdaBoost, train and build models with a feature subset using repeated K-fold cv. • Produce a variable importance score*
9) *Train and build models using the feature subsets generated by Feature Importance with repeated K-fold cv and the ML classifiers RF, GB, and AdaBoost.*
10) *Using test data, evaluate the performance of all models.*
11) *Predict the best feature subset and classification strategy.*
12) *End*

2.2. Data

The required data is imported from HCAHPS website. It is an EHR collected from differenttowns in Australia for predicting the best Home Health Care Services. These agencies offer various services such as Home health services [HHS], Medical social service [MSS], Speech pathology [SP], Occupational therapy [OT], Physical therapy [PT], and Nursing care [NC]. Patients are responding to various metrics assigned to them. The study intends to analyze the top metric for patients in valuing health care services based on the given metrics. A total of 12165 responses were collected, with the inclusion criteria for the research being those only agencies that provide comprehensive services have been recognized. The agency that does not offer the above services was omitted from the analysis. After these conditions were removed, 11,365 responses were taken for analysis. Table 1 displays the number of data collected for research purposes based on the research's inclusion and exclusion criteria.

Table 1. Total number of respondents

Total Response	Offers all the services	Agency that does not offer all the services
12165	11365	800

2.3. Data Pre-Processing

The chosen data set is swapped by median values in the column. Because the median is themiddlemost value, it is the best imputation method for replacing missing values when thereare outliers. The data record contains a total of 11,365 instances after preprocessing. It is further divided into training and test data sets for model development and performance evaluation.

2.4. Feature Selection

By bringing down the number of features in the dataset, feature selection plays an essentialrole in obtaining the best model prediction with reasonable accuracy. It is the technique by which an algorithm intuitively looks for the top features in the given dataset. The feature data is used in the ranking-based feature selection to determine the rank. A surge in data size will result in an inevitable growth in the computational budget. Overfitting data will reduce model efficiency if inappropriate features exist in the data set [10]. Feature selection is one of the methods used to minimize the number of ineffective features and aid in developing an economic model. From 14 columns and 12,165 rows were present in the original data set during data cleaning. After data cleaning, it is condensed to 14 columns.

Again, considering the study's inclusion criteria, the columns are cut down to 8 columns and 11365 rows by leaving missing values greater than the threshold value set. We used the filter method to select the best feature, it is obtained with a variance threshold of 0.7 and univariant, bivariant analysis along with heap map is used in choosing the best feature. To attain the expected outcome, these feature selection is used, to obtain the model classification and evaluation.

3. Model Building and Classification

To classify the best HHCS from the available agencies ML classification algorithmare used on the datasets to predict the accuracy score based on the feature selected. The prevalent classification algorithm like RF, GB, and AdaBoost are used to train and validate the dataset. The obtained results are discussed in the results and discussion section in detail.

3.1.1. Feature Importance

It is a technique to assign a score to input features to predict the target value. givea clear insight on the data, model and also help reduce the input features. The feature importance is considered based on the model predicted using ML classification algorithm

3.1.2. Cross validation

Various ML classification algorithms are run on the dataset to find the best featurefor classifying the HHCS based on the features chosen. As acquainted, over-fitting may occur when training a small data set, K-fold cross-validation is implemented to train the model and determine the best model using many classifiers. It is a model validation method in which samples are split into two data subsets, one to train the model and another to validate the model. Since we employed 10-fold validation, the data subsets are divided into ten. Nine data subsets were used to train the model, with the remaining one used to validatethe model. The overall result of this validation method is an average of ten models. The models are built using the chosen features and the ML classification algorithms RF, GB, and AdaBoost. The model results are discussed in the result section, which includes a full analysis report.

3.1.3. **Performance Evaluation**:

The models' performance is determined using an ML classification algorithm, and the results are validated based on accuracy. Initially, the algorithms like RF, GB and AdaBoost are evaluated using training data with the chosen features by filter method. Thenthe model generated through those features is used to evaluate the feature weigh the feature importance, and the best features from these model accuracies are predicted from the selected model. Based on these results, the predicted accuracy model is built, and the model’s evaluated using ROC curve.

3.2. Findings and Discussion

The results are discussed sumptuously in this part based on our proposed methodology. After data cleansing, data preprocessing is accomplished to find the best function based on the several factors. In this section, we discuss our research work findingsbased on the proposed methodology.

3.2.1. Feature Subset

Feature selection is identifying the best out of available feature and omitting the redundantand irrelevant ones in our classification approach. In classification, the features that do nothave relevant information will perform less accuracy and lead to wrong predicting. Various methods can be used in identify the best feature. In our implementation we used methods like correlation matrix and omitting the columns with the threshold greater than 0.7. Some features show the distribution to be straightforward and, in other cases, negatively skewed. This indicates that there is some interdependence among these features related to outcome. We can check thecorrelation matrix\heat map to get more information related to features for better understanding.

4. Machine Learning classifiers

Logistic Regression

Logistic Regression is used to solve the binary classification problem, with the resultant inthe models being either 0 or 1. The model's accuracy achieved is 98.01, and the model performance recall score is 98%, which indicates that the chosen feature accurately predictspatient satisfaction toward various HHCS.

Random Forest

The decision tree is the concept behind the RF classifier; a hierarchical structure built on features. The decision tree node is divided based on the subset of features that are related with it. It is a set of decision trees constructed from various dataset samples The nodes aredivided according to the Gini index of the selected feature subset [17].

Gradient Boost

It is a machine learning classifier algorithm in which multiple weak learners are trainedto improve each other by producing a good result. At each stage, a new tree is formed by correcting the errors created by previous trees. The obtained result is derived from the best machine learning algorithm in healthcare

Ada Boost

AdaBoost is a well-known classifier in machine learning. It employs the boosting techniqueand combines the weak classifier with the majority voting scheme. The higher the weights in the voting scheme of the final classifier, the higher the accuracy obtained during training. The weights obtained from the recent study is one of the classifiers that outperforms many ML classifiers.

Table 2: Performance of the classifiers

S.No	Classifier	Accuracy
1	Logistic Regression	98.01
2	Random Forest	99.86
3	Gradient Boosting	99.82
4	Ada Boost	86.86

Table 1 shows the performance of the proposed feature selection with various ML classifiers. It is found that the ensemble learning based classifier random forest gives the better performance.

Feature Importance

Following training and optimization of the various ML models, the feature importance scores for each feature chosen were calculated using the model's RF, GB, and ADB. Figure

4.9 explains the ranking on models' best feature importance. The y-axis denotes feature variables, while the x-axis describes feature scores. Based on the score, the significant features Professional, Communication , Safety and Overall Care in patient satisfaction arefound. Out of this , Professional treatment received the highest score, and the safety rendered received the next highest score in determining patient satisfaction for HHCS.

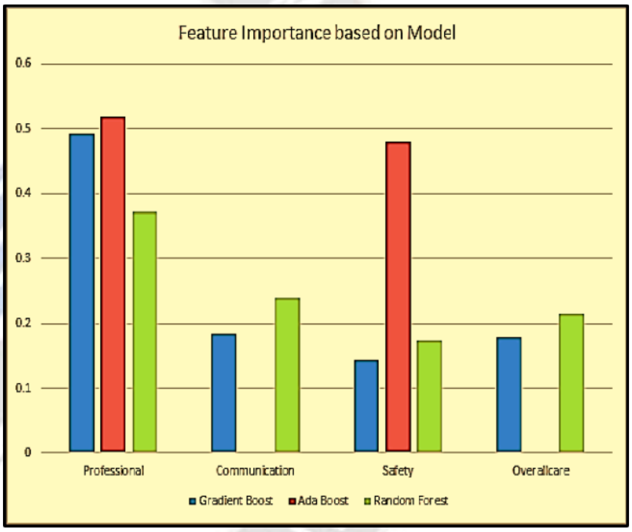

Fig.2 Feature Importance based on model

5. Conclusion

In this work a the novel feature selection method for identifying the key feature in Home Health Care Services using patient satisfaction data is proposed. Following training and optimization of the various ML models, the feature importance scores for each feature chosen were calculated using the model's RF, GB, and ADB. Based on the score, the significant features Professional, Communication , Safety and Overall Care in patient satisfaction are found. Out of this , Professional treatment received the highest score, and the safety rendered received the next highest score in determining patient satisfaction for HHCS. It is found the Boosting methods perform well when considered with other ML models in ranking a feature. The RF model scored 99.868 and the GB model scored 99.824 when compared to earlier studies.

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

References

[1] Ms. Aarati Mahadik et.al., (2016), "Aspect Based Opinion Mining for Identifying customer Preferences", Int. Journal of Engineering Research and Applications, Vol. 6, Issue 2. 127

[2] Mukesh Rawat et al.,. (2016)," A new alley in Opinion Mining using Senti Audio Visual Algorithm",Int. Journal of Engineering Research and Applications, Vol. 6, Iss.2, pp.01-09.

[3] Nasukawa et.al., (2003),"Sentiment analysis:Capturing favorability using natural language processing", KCAP-03.

[4] O'Toole RV, et.al., (2008), LEAP Study Group. Determinants of patient satisfaction after severe lower-extremity injuries. J Bone Joint SurgAm, Vol.90, Iss.6,pp.1206-11.

[5] P. Baranikumar et.al., (2016), "Feature extraction of opinion mining using ontology", Int. J. Adv. Comput. Electron. Eng., vol. 1, no. 1, pp. 18-22,

[6] Pak, Alexander et.al.,. (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of LREC. 10.

[7] Palla Pavankumar et al., (2016) "Customer Reviews and Analysis using Opinion Mining Adaptive Algorithm", IJIRCCE, Vol 4, Iss.4.

[8] Paltoglou G, et.al., (2010),study of information retrieval weighting schemes for sentiment analysis. In: Proceedings of the 48th annual meeting of the association for computational linguistics (ACL "10). pp. 1386–1395.

[9] Panchal, Dnyaneshwar et.al.,(2020). "Sentiment Analysis of Healthcare Quality", International Journal of Innovative Technology and Exploring Engineering. 09. pp.1-8.

[10] Pichert, J. W., et.al.,(2008). Using patient complaints to promote patient safety

[11] Poornappriya, T.S., Gopinath, R., Application of Machine Learning Techniques for Improving Learning Disabilities, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 392-402 (2020).

[12] Poornappriya, T.S., Selvi, V., Evolutionary Optimization of Artificial Neural Network for Diagnosing Autism Spectrum Disorder, International Journal of Electrical Engineering and Technology (IJEET), 11(7), 47-61 (2020).

[13] Priyadharshini, D., Poornappriya, T.S., & Gopinath, R., A fuzzy MCDM approach for measuring the business impact of employee selection, International Journal of Management (IJM), 11(7), 1769-1775 (2020).

[14] Poornappriya, T. S., and R. Gopinath. "Segmentation Of Cervical Cancer Lesions: A Comparative Analysis Of Image Processing Algorithms." *Webology (ISSN: 1735-188X)* 18.4 (2021).

[15] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems* 37.5 (2019): 6743-6760.

[16] Durairaj, M., and T. S. Poornappriya. "Why feature selection in data mining is prominent? A survey." *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019*. Springer International Publishing, 2020.