International Journal of Hospital Research 2021, 10(1)
http://ijhr.iums.ac.ir
Research Article

Machine Learning Algorithms for prediction of in-patients satisfaction

Toktam Khatibi[1], Rouhangiz Asadi[2], Mohammad Mehdi Sepehri[1*], Pejman Shadpour[2]

[1]Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[2]Hasheminejad Kidney Center, Hospital Management Research Center, Iran University of Medical Sciences, Tehran, Iran

Abstract

Background and Objective: The health industry is a competitive and lucrative industry that has attracted many investors. Therefore, hospitals must create competitive advantages to stay in the competitive market. Patient satisfaction with the services provided in hospitals is one of the most basic competitive advantages of this industry. Therefore, identifying and analyzing the factors affecting the increase of patient satisfaction is an undeniable necessity that has been addressed in this study.

Methods: Because patient satisfaction characteristics used in hospitals may have a hidden relationship with each other, data mining approaches and tools to analyze patient satisfaction according to the questionnaire used We used the hospital. After preparing the data, the characteristics mentioned in the questionnaire for patients, classification models were applied to the collected and cleared data, and with the feature selection methods, effective characteristics Patients were identified and analyzed for satisfaction or dissatisfaction.

Results: Based on the findings of the present study, it can be concluded that the factors of patient mentality of the physician's expertise and skill, appropriate and patient behavior of the physician and food quality (hoteling) respectively have a higher chance of increasing patient satisfaction with Establish services provided in the hospital.

Conclusion: Comparing the approach used in this study with other studies showed that due to the hidden effects of variables on each other and the relatively large number of variables studied, one of the best options for analyzing patient satisfaction questionnaire data, Use of data mining tools and approaches

Keyword: Machine Learning Algorithms, Patient satisfaction, Data Mining, Clustering, Feature selection

Background and Objective

Hasheminejad kidney center, based on its strategic mission, pays special attention to discovering and meeting patients' expectations as the main axis of the hospital. Based on the model of European Customer Satisfaction Index (ESCI), this center reviews, discovers and measures customer satisfaction every year[1]. In this study, we seek to analyze the satisfaction of patients admitted to Hasheminejad Hospital with the help of data mining. According to the above explanations, the benefits and necessity of patient satisfaction analysis in the hospital are undeniable.

In many previous studies, statistical analysis and statistical tests have been used to measure patient satisfaction[2,3, 4-12]. In recent years, researchers have considered the use of machine learning techniques to analyze patient satisfaction to make structural changes in the provision of medical services[13-15].Galatas et al. analyzed patient satisfaction using data mining techniques In their study, they showed that the results of data mining analysis are significantly related to the results of statistical analysis[13].

*Corresponding Author: Mohammad Mehdi Sepehri
Email: mehdi.sepehri@gmail.com

Due to the large dimensions of the questionnaire used in the present study and the large number of hospitalized patients, statistical approaches in analyzing these data will fail, machine learning approaches are proposed to analyze customer satisfaction measurement data in this project. The reason we will use machine learning to analyze patient satisfaction is that we will examine several dimensions of characteristics to measure patient satisfaction. Due to the large number of patients and the variety of characteristics under study (complex data), machine learning provides us with appropriate tools to be able to use them to analyze data with multiple dimensions.

Method

Figure 1 shows the stages of research methodology (research framework).

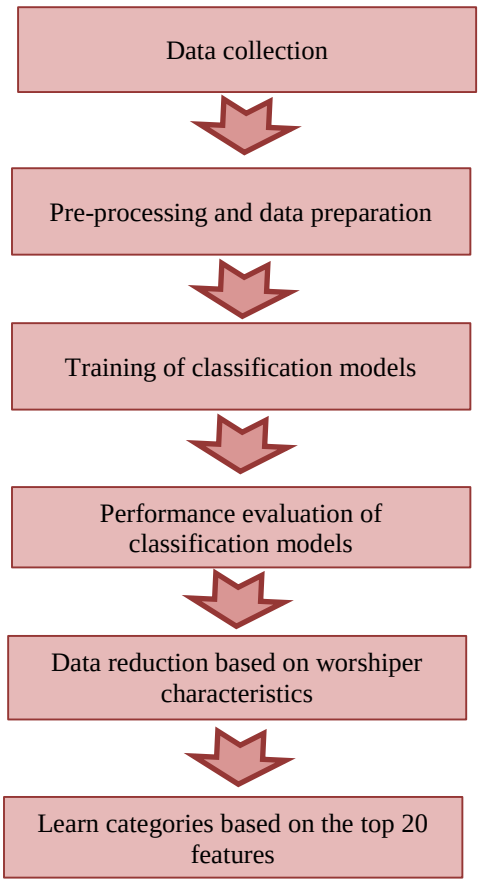


Figure 1. Research performance framework

According to Figure 1, the research framework includes the steps of data collection, data preprocessing, classification model training, classification model performance evaluation, data reduction based on differentiating features, and finally classification training based on the top 20 features that is explained in detail.

Data collection

The Self-made questionnaire[1] was used in order to collect data in this study. For this purpose, the questionnaires completed the questionnaire in person to person with questions and answers from patients. For this purpose, patients who had an acceptable level of alertness to respond or their companions were selected to participate in the survey. Since the average length of hospital stay of patients in wards (except transplant ward and outpatient surgery ward) is 3 days, to prevent data bias, data were collected from wards 3 days apart. In this way, a special ward was referred once every three days to collect data so that patients were discharged 3 days earlier and participants were not duplicates.

Features

Questionnaire[14] information of 500 in-patients was collected for a period of three months in the form of an Excel file containing 75 columns (features). After deleting the first and last name columns, the data had 74 columns. In many cases, columns 60 to 74, which are related to the supplementary questions, had missing values. Due to the large number of missing values (about 90% of the rows), these columns were removed from the data.

Columns one to 59 of the data are as follows:

- If1 to If10 are personal profile questions.
- N1 to N9 are nursing questions.

- M1 to M7 are medical questions.
- H1 to H8 are hoteling questions.
- R1 to R8 are communication questions.
- IM1 to IM5 are patient mentality questions from the hospital.
- C1 to C1.1 are complaint questions.
- L1 to L4 are loyalty questions.
- Q1 to Q4 are general questions.
- the patient's score to the hospital.

"if4" "N2" "M1" "M2" "M5" "M6" "M7" "H1" "H5" "H6" "H7" "H8" "R3" "R7" "IM2" "L3" "Q3"

On the other hand, all If9 column values were 1 or the missing value. Therefore, this column was also removed. Columns If6.1 and C1.1 also had more than a third of the missing values and were therefore removed.

Pre-processing and data preparation

Usually, the data collected can not be used in the current format and it is necessary to take steps to clear and prepare the data on it. In many cases, the data contains noise, outlier points, missing values, and various intervals for different characteristics. If the data cleaning and preparation steps(data preprocessing) are not correctly performed, the results obtained from clasification models and other machine learning models will not be significant and may even be invalid. Therefore, it is necessary to take the necessary steps to prepare data on it)[16] .

The steps required to prepare the data in this study include identifying and replacing missing values, normalizing numerical characteristics, and finally discretizing the class label column. Each of these steps will be described in detail below.

Identify and replace missing values

Rows with an unspecified amount of points are deleted, because the score determines the class label. Also columns with more than a quarter of the missing value data are removed.After examining different threshold values to remove the columns, the threshold value was selected as a quarter, which led to more accuracy, precision and recall of the classification models in both training data and test data.

On the other hand, columns which all registered values are the same number for thos are removed.

The KnnImputation command in the DMWR software package in R software is used to locate the missing values. In this method, to fill in the missing values, the nearest neighbors of the row with the missing value are identified in terms of other characteristics and the average value of that column in these nearest neighbors is considered as the suggested value for the row with the missing value. Neighbor K = 3 was selected to replace the missing values. The reason for choosing three neighbors is that the noise data is more involved with the choice of K = 1 neighbor, and on the other hand, increasing K leads to the involvement of data in determining the missing value of a data that is less similar to it. Therefore, the best recommended value is K = 3.

Normalization of numerical data

The min-max method is used to normalize numerical data. In this way, the value of the numbers in that column is normalized according to the following equation for each column with numeric values x:

$$normalizedX = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Discrete the class label column

we first plotted the class label column histogram In order to discretize the class label column, which is shown in Figure 2.

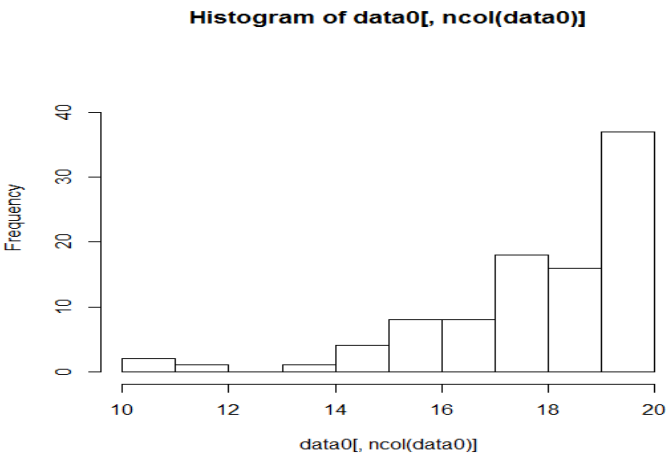


Figure 2 *Histogram of the class label column*

Since more than half of the comments assigned a total score above 18, we considered two satisfaction classes with a score above 18 and a score less than or equal to 18 for discretization, which is about 55.8% of patients in The first class and the others were in the second class. Experts were used to discretize the class label column and the selected intervals were approved by the experts.

Classification model training

In this study, random sampling was performed by placing an equal number of both classes for data sampling, and the data that did not participate in the training sample to build a classification model were considered as a test or test set.

The decision tree classifier was created using this tutorial. Another model used to classificate patients' opinions is the model of support vector machines in the radial kernel. The next model is linear kernel support vector machines (SVM)[17].

The decision tree classifier is a popular, simple, and fast model that works well for classifying linearly separable data. On the other hand, the graphic diagram and association rules extracted from this category can help analyze the relationships between variables and better understand the performance of the model.

Support machine classifier is one of the strongest classifications that usually has a very good performance in data classification. This category can use different kernels to separate different types of data patterns. For example, for linearly separable data, linear kernels are used, and for linearly separable data, kernels such as radial, polynomial, and sigoid kernels are used. Usually the radial kernel performs better than other kernels for linearly integral data.

Evaluate system performance

In order to evaluate the performance of the constructed category, the criteria of accuracy, sensitivity and specificity and F score are used.

The formulas for accuracy, sensitivity and specificity are shown below:

$$accuracy = \frac{TP + TN}{n}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

In the above relations, TP is the number of true positive examples, TN is the number of true negative examples, FP is the number of positive positive examples and FN is the number of negative negative examples. n is also the total number of examples.

$$F - Measure = 2\frac{Sensitivity * Specificity}{Sensitivity + Specificity}$$

Reduce data based on selection of differentiating features

Because the performance of the classification models was sometimes not very favorable, we decided to use the feature selection methods and identify the distinguishing features and then study the classification model only on the superior set of distinguishing features to see if improvements in the performance of classification models can be provided to measure patient satisfaction. For this purpose, among the feature selection methods, the random forest method was used.

After applying the random forest method by constructing 100 trees on the educational data, the top twenty features were identified in

terms of average reduction of error squares and reduction of node purity. For this purpose, twenty superior properties were extracted to reduce the mean error squared and twenty superior properties were extracted to reduce the impurity of the nodes.

Characteristics that were among the top twenty characteristics in both feature selection methods with random forest in terms of average reduction of error squared and reduction of node purity were also identified.

When selecting and reducing attributes is done with filtering methods such as correlation analysis, selecting and reducing the number of attributes will be one of the steps of data preprocessing. However, since this operation can be performed based on the results of the importance of variables after training the categories, it can not be in the pre-processing and data preparation stage.

Clasifier training with only the top twenty distinguishing features

At this stage, decision tree categories and support vector machines, which are identified using only the top twenty features identified by one of the two methods, are evaluated and performance criteria are evaluated.

The decision tree was applied to the top twenty properties to reduce the mean of the error squares. Radial kernel support vector machines were also applied to the top twenty characteristics to reduce the mean error squared. Linear kernel support vector machines were also applied to the top twenty characteristics to reduce the average error squares. The decision tree was applied to the top twenty properties to reduce the purity of the nodes.

Radial kernel support vector machines were applied as another classifier to the top twenty characteristics to reduce node impurity. Support vector machines with linear kernels were also applied to the top twenty characteristics to reduce node impurity.

Finally, it is necessary to evaluate the performance of classification models using the top twenty differentiating features in one of two ways: reducing the average error squared or reducing the imperfection of nodes. And compare which category can achieve the highest level of performance.

Results

The decision tree clasifier was created using a training sample, which is summarized in Figure 3 and Figure 4.

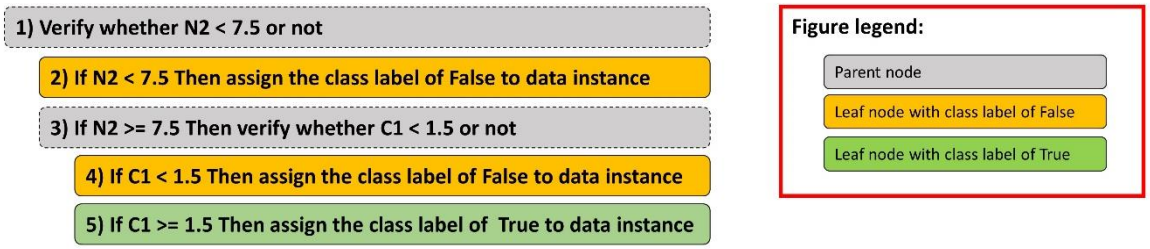
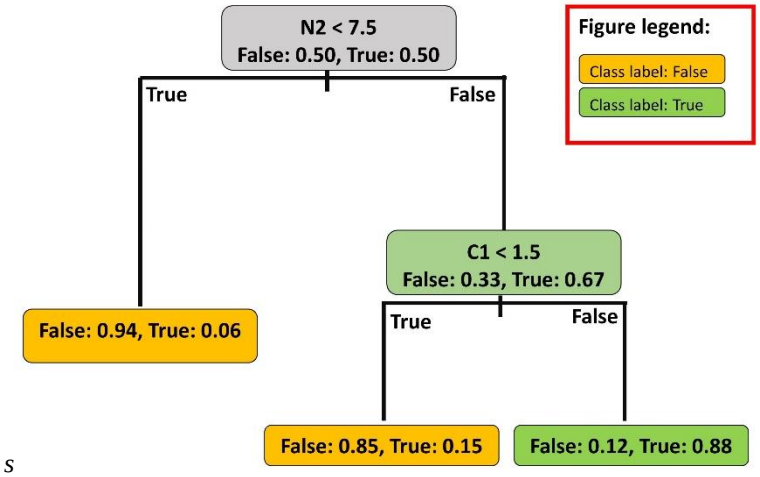


Figure 3. Summary of the decision tree made to categorize patient opinion



s

Figure 4. Decision tree made to clasify patients' opinions

Evaluate system performance

Table 2 summarizes the classification models by performance evaluation indicators.

Another model used to categorize patients' opinions is the model of support vector machines in the radial kernel. The next model is the model of support vector machines with linear kernels.

Table 2. Performance evaluation of different classification models to measure patient satisfaction

Clasifier model	Accuracy	sensitivity	Being exclusive	F score
decision tree	69.8	57.14	78.12	66.00
Backup vector machine with radial kernel	69.6	88.9	57.1	69.54
Backup vector machine with linear kernel	65.3	61.1	67.8	64.27

According to Table 2, all classifiers performed better than the random classification method with an accuracy of 0.5 on the test data, but their performance was not very strong.

Reduce data based on selection of differentiating features

After applying the random forest method by constructing 100 trees on the educational data, the top twenty features were identified in terms of average reduction of error squares and reduction of node purity. Figure 5 illustrates the top twenty features to reduce the mean error squared, and Figure 6 illustrates the top twenty features to reduce the impurity of the nodes.

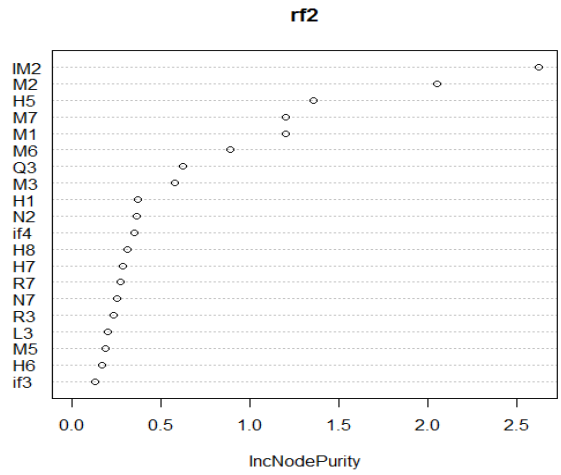


Figure 6. 20 Superior distinguishing feature in terms of purity of tree nodes

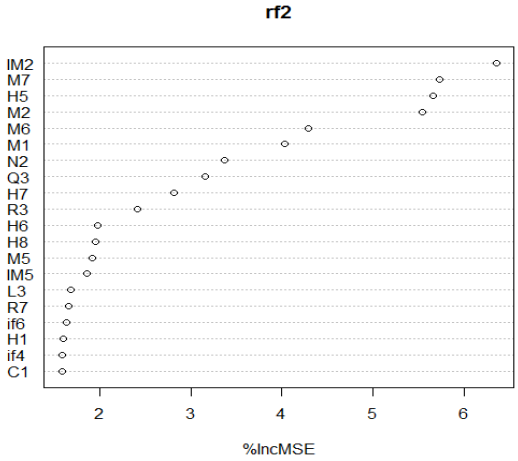


Figure 5. Top 20 Distinctive Characteristics in terms of Mean Squared Error

The characteristics that were among the top twenty characteristics in both methods of feature selection with random forest in terms of average reduction of error squares and reduction of node purity are:

"if4" "N2" "M1" "M2" "M5" "M6" "M7" "H1" "H5" "H6" "H7" "H8" "R3" "R7" "IM2" "L3" "Q3

clasifier training with only the top twenty distinguishing features

At this stage, decision tree clasifiers and support vector machines are taught using only the top 20 features identified by one of the two methods, and performance metrics are evaluated.

Figure 7 summarizes the decision tree applied to the top twenty properties to reduce the mean error squared."

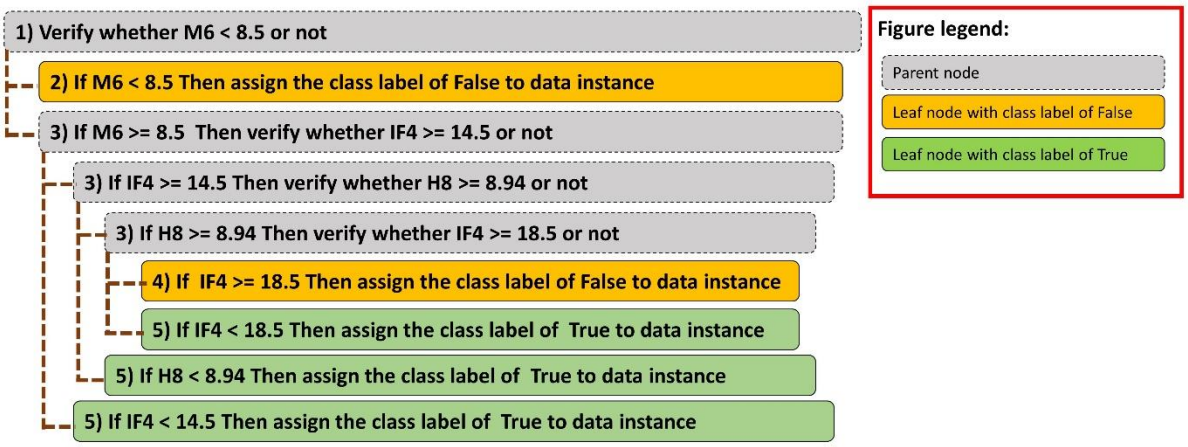


Figure 7. Summary of the decision tree on the top twenty features to reduce the average error squared

Figure 8 shows the decision tree obtained with these twenty top features.

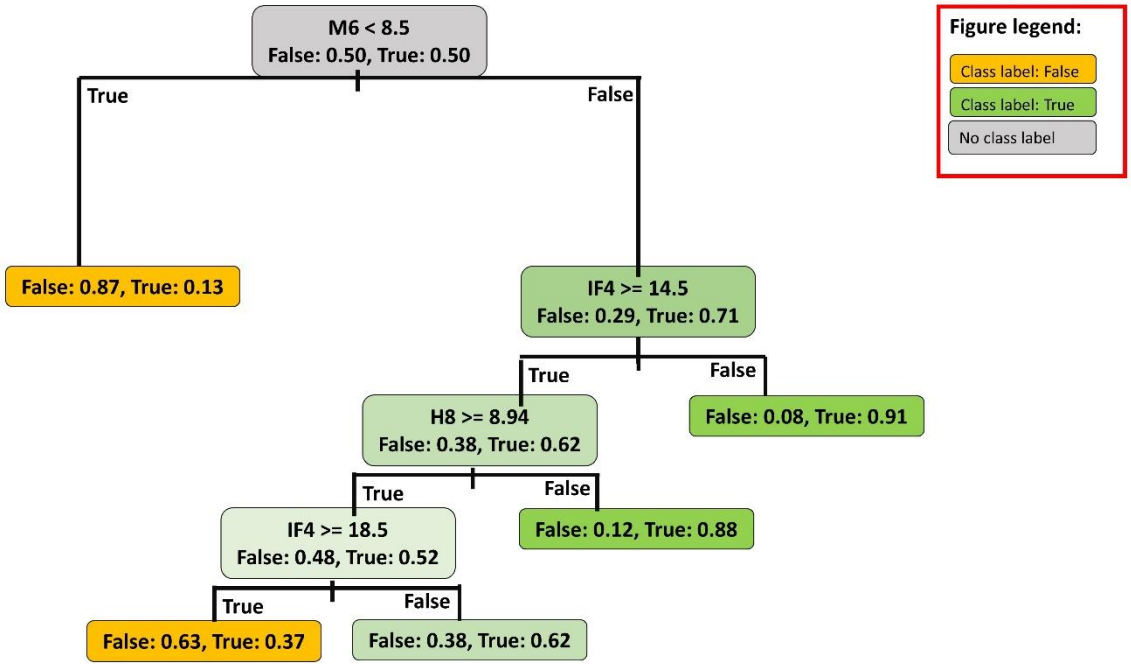
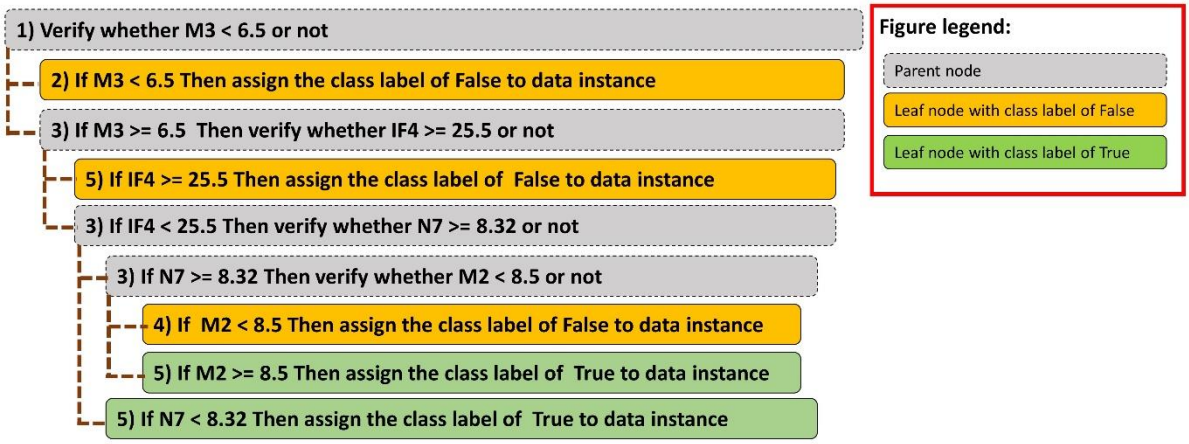


Figure 8. The decision tree obtained from the top twenty properties in order to reduce the average error squared

Radial kernel support vector machines were also applied to the top twenty characteristics to reduce the mean error squared. Linear kernel support vector machines were also applied to the top twenty characteristics to reduce the average error squares.

Figure 9 summarizes the decision tree applied to the top twenty properties to reduce the impurity of the nodes.



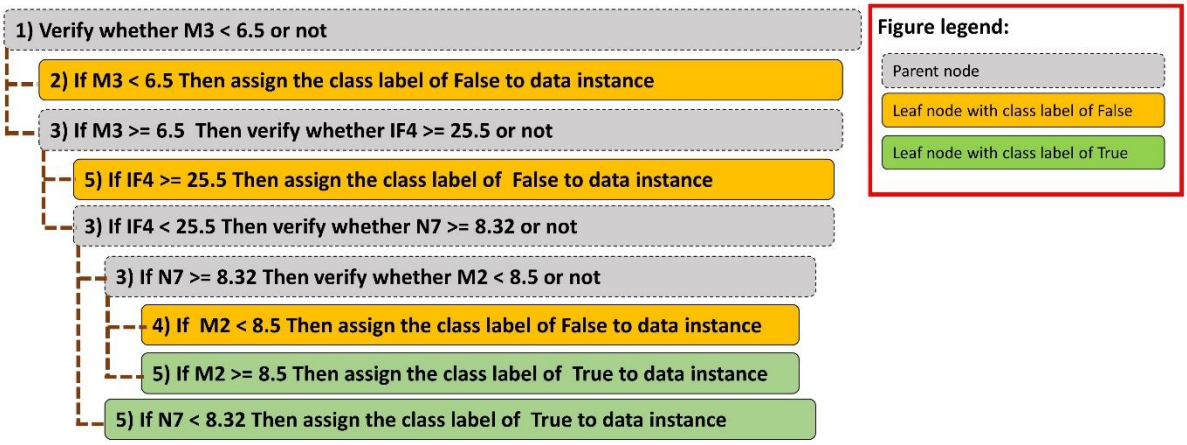


Figure 9. Summary of the decision tree on the top twenty properties to reduce the impurity of the nodes

Figure 10 shows the decision tree obtained with these twenty top features.

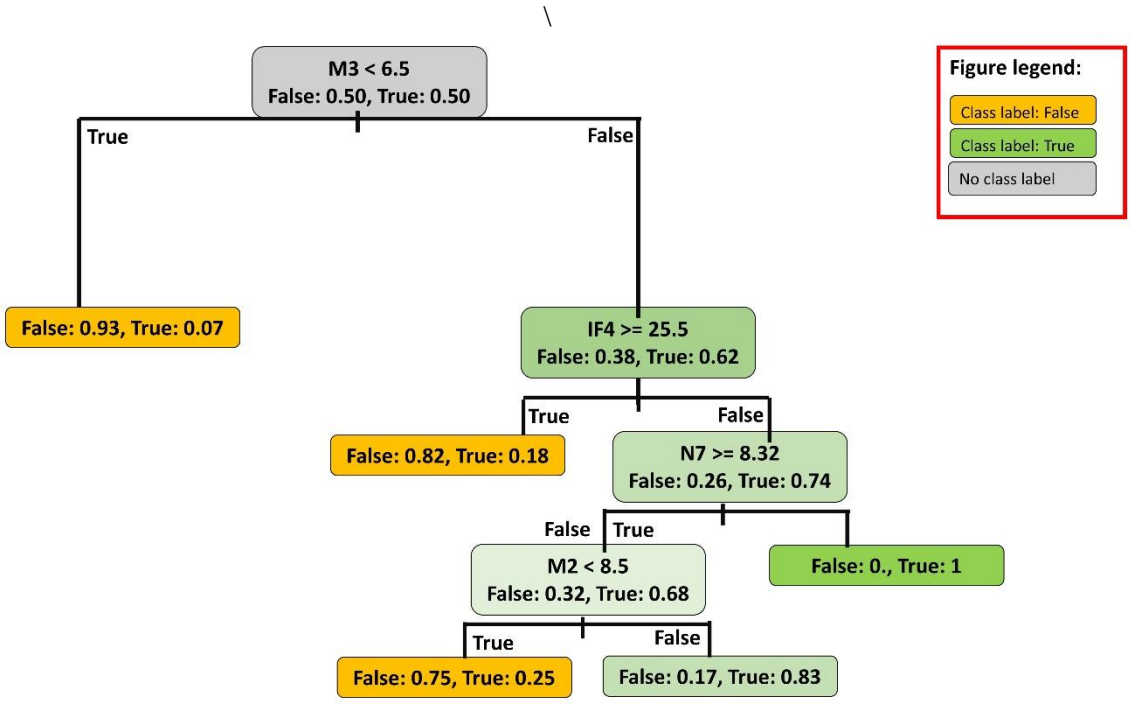


Figure 10. Decision tree from the top twenty properties to reduce the impurity of nodes

Radial kernel support vector machines were also applied to the top twenty characteristics to reduce node impurity.

Support vector machines with linear kernels were also applied to the top twenty characteristics to reduce node impurity.

Table 3 summarizes the performance evaluation of classification models using the top twenty distinguishing features in one of two ways: reducing the mean error squared or reducing the imperfection of the nodes.

Table 3. Evaluate the performance of category models using them based only on the top twenty features

Features used	Clasifier model	Accuracy	sensitivity	Being exclusive	F score
Reduce MSE	decision tree	0.6	0.68	0.55	0.61
Reduce MSE	Support vector machine with radial kernel	0.68	0.84	0.58	0.69
Reduce MSE	Support vector machine with radial kernel	0.62	0.53	0.68	0.59
Reduce impurity	decision tree	0.62	0.5	0.7	0.58
Reduce impurity	Support vector machine with radial kernel	0.8	0.8	0.8	0.8
Reduce impurity	Support vector machine with radial kernel	0.58	0.35	0.73	0.47

According to the results of Table 3 and its comparison with Table 2, we conclude that the best performance of that category is the support vector machines with radial kernel, which has been applied to the top twenty features to reduce the impurity of the nodes. Other clasifiers do not show a significant difference in the case that applies to all features compared to the case that applies to the top twenty features.

Conclusion

One of the main goals of the hospital is customer orientation and obtaining high patient satisfaction. Hasheminejad subspecialty center during the past years, at regular intervals and based on a systematic method has surveyed patients in different wards of the hospital. In each period, based on the results of the survey, hospital policies and strategies are adopted. The analysis of patients' opinions in the hospital so far has been based on statistical results (average satisfaction) in different wards. In this study, we examined several dimensions of characteristics to analyze patient satisfaction. Due to the large number of patients and the diversity of the studied characteristics (complex data), data mining is a suitable tool for data analysis with large dimensions, which was used in this study. Based on the findings of the present study, it can be concluded that the following factors create a higher chance of patients' satisfaction with the services provided in the hospital, respectively.

{Patient mentality> personal characteristics > Hoteling}….→ . Patient satisfaction.

In the above model, the patient mentality refers to the good and appropriate relationship between the physician and the patient, and the personal characteristics of the physician are returned, and the quality of food in the field of hoteling is effective in patient satisfaction. As can be seen, the credibility and appropriate behavior of the physician has a great impact on patient satisfaction.

Data mining reveals hidden data and information between survey results, the results of this study can help hospital managers and officials to analyze patients' behavior in choosing a hospital, their loyalty to the hospital, and their mentality of the hospital. And guide them in future planning for patients. It will also help managers plan for patients who are more loyal to the hospital by categorizing patients.

In our previous research, the use of data mining approaches to analyze patient satisfaction has received less attention. However, due to the hidden effects of variables on each other and the relatively large number of variables studied, one of the best options for analyzing patient

satisfaction questionnaire data is to use data mining tools and approaches. Therefore, this case is considered as a kind of innovation of this research.

Since not every questionnaire can be comprehensive and complete and show all the cases considered by patients in measuring their satisfaction with the services they have received. On the other hand, filling out the questionnaire is sometimes difficult and long for patients and makes them and their companions bored. This can reduce the accuracy of patients in filling out the questionnaire and cast doubt on the validity of the research results on the data of completed questionnaires.

Therefore, it is suggested that patients' opinions be recorded briefly in the form of text or even speech, and then using text processing and text mining systems as well as speech to text conversion to try to analyze the comments and measure their satisfaction. On the other hand, the system can also analyze patients' suggestions.

Acknowlge

The authors consider it necessary to appreciate the cooperation and efforts of the Quality Manager of Hasheminejad Hospital and also the cooperation of the Hospital Management Research Center in advancing this research.

Competing interest:

The authors declare no competing interest

Authors contributions:

The Authors have the same contributions in this study.

Refrences

1. Semnani F, Ansar F, Asadi R. Designing the Hybrid Model of EFQM- ECSI for Evaluating the Satisfaction of Patients: Case of Hasheminejad Hospital. Int J Hosp Res. 2017;6(3).
2. D.C.Ferreira[a]R.C.Marques[a]A.M.Nunes[a]J.R.Figueira. Patients' satisfaction: The medical appointments valence in Portuguese public hospitals. Omega. Volume 80, October 2018, Pages 58-76
3. Assila Anis Asnawi, Zainudin Awang, Asyraf Afthanorhan* , Mahadzirah Mohamad and Fazida Karim. The influence of hospital image and service quality on patients' satisfaction and loyalty. Management Science Letters 9 (2019) 911–920.
4. Guiahi, M., Sheeder, J. & Teal, S. 2014. Are women aware of religious restrictions on reproductive health at Catholic hospitals? A survey of women's expectations and preferences for family planning care. *Contraception,* 90**,** 429-434.
5. Baiardini, I., Puggioni, F., Menoni, S., Boot, J. D., Diamant, Z., Braido, F. & Canonica, G. W. 2013. Patient knowledge, perceptions, expectations and satisfaction on allergen-specific immunotherapy: a survey. *Respiratory medicine,* 107**,** 361-367.
6. Krishnasamy, M., Ugalde, A., Carey, M., Duffy, M. & Dryden, T. 2011. Patient expectations and preferences for follow-up after treatment for lung cancer: a pilot study. *European Journal of Oncology Nursing,* 15**,** 221-225.
7. Wathoni, N. & Rahayu, S. A. 2014. A survey of consumer expectation in community pharmacies in Bandung, Indonesia. *Journal of Applied Pharmaceutical Science,* 4**,** 84
8. Péntek, M., Brodszky, V., Gulácsi, Á. L., Hajdú, O., Exel, J., Brouwer ,W. & Gulácsi, L. 2014. Subjective expectations regarding length and health-related quality of life in Hungary: results from an empirical investigation. *Health Expectations,* 17**,** 696-709.
9. Halpert, A., Dalton, C. B., Palsson, O., Morris, C., Hu, Y., Bangdiwala, S., Hankins, J., Norton, N. & Drossman, D. A. 2010. Irritable bowel syndrome patients' ideal expectations and recent experiences with healthcare providers: a national survey. *Digestive diseases and sciences,* 55**,** 375-383.
10. Linde, K., Witt, C. M., Streng, A., Weidenhammer, W., Wagenpfeil, S., Brinkhaus, B., Willich, S. N. & Melchart, D. 2007. The impact of patient expectations on outcomes in four randomized controlled

trials of acupuncture in patients with chronic pain. *Pain*, 128**,** 264-271.
11. Chang, W.-J. & Chang, Y.-H. 2013. Patient satisfaction analysis: Identifying key drivers and enhancing service quality of dental care. *Journal of Dental Sciences*, 8**,** 239-247.
12. Vijaya Sunder MORCID Icon,Sanjay Mahalingam &Sai Nikhil Krishna M .Improving patients' satisfaction in a mobile hospital using Lean Six Sigma – a design-thinking intervention. Journal of Production Planning & Control , The Management of Operations . Volume 31, 2020 - Issue 6. Pages 512-526.
13. Georgios Galatas, Dimitrios Zikos, Fillia Makedon. Application of Data Mining Techniques to Determine Patient Satisfaction. Conference Paper · May 2013. DOI: 10.1145/2504335.2504379.
14. Deirdre Mylod, Dennis Kaldenberg. Data Mining Techniques for Patient Satisfaction Data in Home Care Settings. 2000Home Health Care Management & Practice 12(6):18-29. DOI: 10.1177/108482230001200607.
15. Masumi Okuda; Akira Yasuda; Shusaku Tsumoto. A Data Mining Approach on the Structure of Patient Satisfaction in HCAHPS Databases. Published in: 2016 IEEE International Conference on Healthcare Informatics (ICHI).
16. Han J., Kamber M., Pei J. , (2012), Data mining: Concepts and Techniques, Morgan Kauffmann.
17. Amir Ahmad. Decision tree ensembles based on kernel features. 2014. Applied Intelligence volume 41, pages855–869(2014).

Please cite this article as:

Toktam Khatibi, Rouhangiz Asadi, Mohammad Mehdi Sepehri ,Pejman Shadpour. Machine Learning Algorithms for the prediction of the in-patients satisfaction Int J Hosp Res. 2021;10 (1).