Thirty-day Re-Hospitalization Rate Prediction of Diabetic Patients

Dong-Her Shih[1], Feng-Chuan Huang[1], Cai-Ling Weng[1], Po-Yuan Shih[2], David C. Yen[3]

[1] Dept. of Information Management, National Yunlin University of Science and Technology, Taiwan
[2] Department of Finance, National Yunlin University of Science and Technology, Taiwan
[3] Jesse H. Jones School of Business, Texas Southern University, USA
shihdh@yuntech.edu.tw, d10323004@yuntech.edu.tw, boon0724oao@gmail.com,
D10424003@gemail.yuntech.edu.tw, David.Yen@tsu.edu

Abstract

Diabetes is a serious global health problem, and re-hospitalization is usually associated with increased mortality and excessive medical burden. With the increasing cost of diabetes to the health care system, re-hospitalization is recommended as a goal to reduce health care costs. This paper aims to use data mining technology to accurately predict the 30-day re-hospitalization of diabetic patients. We use the data set from UCI machine learning repository, preprocessing, use feature reduction method to find out the classification results of re-hospitalization, and then use frequent set and Apriori algorithm to find the association rules between diabetes mellitus patients and re-hospitalization related variables. The experimental results show that the recursive feature reduction method is effective in combined with SVM can get a better prediction accuracy.

Keywords: Re-hospitalization, Diabetes, Features reduction, Data mining, Association rules

1 Introduction

According to the statistics of the National Health Agency of the Ministry of Health and Welfare [1], diabetes mellitus ranks the fifth among the top ten causes of death in China, in 2017, there are about 425 million adults suffering from diabetes in the world, and the global prevalence rate of diabetes has reached up to 8.8%. It is estimated that the total number of diabetes patients will increase to 629 million by 2045, and the medical expenditure for diabetes is estimated to reach 776 billion US dollars per year [2]. Diabetes has become a serious global health problem.

Re-hospitalization is defined as a short period of time (usually within 30 days) occurred after the first hospitalization, which may be due to expected or unexpected reasons. Re-hospitalization is usually closely related to increased mortality and excessive medical burden [3]. According to clinical experience, it is difficult to identify the patients who may be re-hospitalized in the next 30 days after the discharge [4], which is attributed to the complex aetiology of re-hospitalization. Therefore, hospitals often need to know among the patients, who are at higher risk for re-hospitalization before discharge [5]. American Diabetes Association [6-8] pointed out that most of the expenses of diabetic patients are from hospitalization. Compared with those without diabetes, the average length of stay and cost of hospitalization are higher; at least 30% of these hospitalized diabetic patients would stay in the hospital again, accounting for more than 50% of the total hospitalization and medical expenses [9-10], with the increasing cost of the health care system on diabetes, re-hospitalization is considered as an indicator of hospital performance and efficiency, which is regarded as a goal to reduce health care costs.

There have been many related studies for evaluating the re-hospitalization rate of patients, most of which use the statistical methods of existing data [11]. The most commonly used methods to predict re-hospitalization are Stepwise Logistic regression [12-13], and multiple Logistic regression [14-15]. Most of the past related studies used statistical methods as the main analysis method, however, believe that in general, statistical models cannot provide accurate predictions of re-hospitalization for specific patients [16]. Some statistical models with poor performance have presented an accuracy rate of less than 50% on average and rarely exceed 70%. Other methods can achieve better results for accurate prediction, such as data mining and machine learning that tend to bring better contribution [17].

With the rise of health awareness, more and more people are beginning to pay attention to their own health management. This research tries to use data mining technology to accurately predict the possibility of re-hospitalization of diabetic patients in 30 days, improving the poor prediction accuracy and use the

*Corresponding Author: Dong-Her Shih; E-mail: shihdh@yuntech.edu.tw
DOI: 10.3966/160792642020122107021

association rules in data mining technology and the Cart decision tree to find out the association between diabetic patients and re-hospitalization related variables and the risk group of re-hospitalization. Finally, comparing the obtained research results with the research of other scholars to offer to the hospital or medical staff for reference.

2 Methodology

The purpose of this research is to predict whether diabetic patients will be hospitalized again in 30 days through data mining technology. In the pre-processing part, three different methods are used for variable reduction comparison, and then four data mining techniques are used to select the best predictive model. At the same time, the Apriori algorithm and the Cart decision tree are used to find the variables related to diabetic patients and re-hospitalization, and, the association between and the risk group of re-hospitalization.

2.1 Experimental Design

This chapter aims to illustrate the research method, which is shown in Figure 1. First, the data set of re-hospitalized diabetic patients was cleaned and integrated, and then three feature reductions were compared, namely rough set feature reduction, Boruta feature reduction, and recursive feature reduction. After that, the data after the three feature reduction methods were subjected to 10-layer cross-validation to divide the training data and test data. Then the training data was trained to establish a data classification model. This study uses SVM, Naïve Bayes, and C45 The four classification methods of decision tree and logistic regression are used to establish a classification model to evaluate the accuracy of the model prediction. Finally, Apriori algorithm and CART decision tree are used to analyze the results of association rules and the result of the re-hospitalization risk to evaluate the accuracy of the model, and use Apriori algorithm and CART decision tree to find the association rules and high and low risk groups with re-hospitalization.

2.2 Materials

This research experiment uses the diabetes data set of 130 US hospitals from 1991 to 2008 in the UCI Machine Learning Repository (Diabetes 130-US hospitals for years 1999-2008), with a total of about 100,000 data and 50 attributes [18]. The selected hospital information must meet the following conditions:

(1) An inpatient (a hospital admission).

(2) Any kind of diabetes can be diagnosed and managed under its medical care system.

(3) The length of stay was at least 1 day and 14 days at the maximum.

(4) Can perform lab tests of glucose.

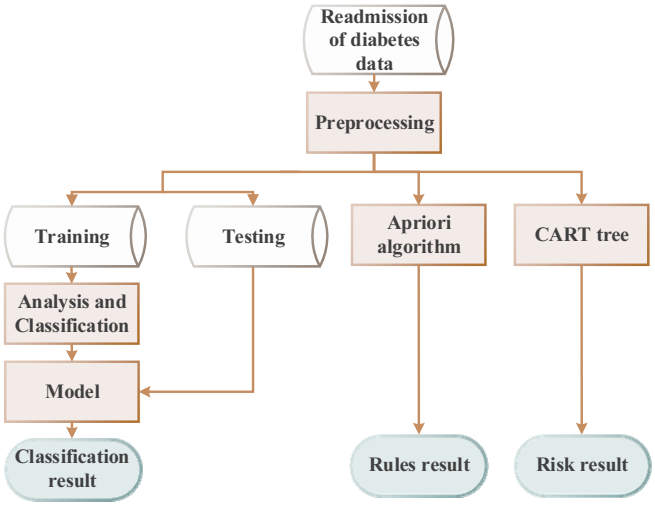


Figure 1. Research flow chart

(5) Giving prescription of medications.

2.3 Performance Index

The classification matrix, also known as the "confusion matrix", is an important tool for evaluating prediction results. By examining the number and percentage of each data grid of the classification matrix, you can quickly check the accuracy of the model prediction. Classification can be divided into True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). In predicting patients' 30-day re-hospitalization, they can be divided into two categories, 30-day re-hospitalization and non-30-day re-hospitalization.

The classification matrix can evaluate the correctness of the classification model in the model classification decision, shown in Table 1 [19]:

Table 1. Classification matrix

		Predicted outcome	
		re-hospitalization	Unplanned re-hospitalization
True condition	Re-hospitalization	TP	TN
	Unexpected re-hospitalization	FP	FN

- True Positive (TP): Diagnosed that the patient would be hospitalized again in 30 days, and the patient did get hospitalized again after 30 days
- False Positive (FP): The patient was predicted to get re-hospitalized in 30 days after discharge, but did not eventually.
- True Negative (TN): predicted that the patient would not get hospitalized again in 30 days, but he/she did eventually.
- False Negative (FN): Predicted that the patient would not get hospitalized again in 30 days, and he /she did not eventually.

The analysis performance can calculate the

following four evaluation indicators: Accuracy, Precision, Sensitivity and Specificity. They will all be used as the evaluation indicators in this study. The higher the value is, the better the prediction performance can be:

- The accuracy rate is the correct classification rate, such as equation (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(1)}$$

- Accuracy, such as equation (2):

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(2)}$$

- Sensitivity is also called recall rate, which means that the classifier recognizes re-hospitalized patients, and the predictive result is also the ability of re-hospitalization [20], such as equation (3):

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(3)}$$

- Specificity indicates the ability of the classifier to identify non-re-hospitalized patients, and the prediction result is also the ability of non-re-hospitalized patients, such as equation (4):

$$Specificity = \frac{TN}{TN + FP} \quad \textbf{(4)}$$

In statistical methods, Odds ratio is a common indicator to measure the correlation between exposure risk and the occurrence of an event. In the medical field, Odds ratio is more likely be used than relative risk because Odds ratio can be used in prospective studies as well as retrospective studies, and its statistical properties are relatively lower is good.

The independent chi-square test is also called the Pearson chi-square test, and can also be referred to as the chi-square test for short. In medical research, the chi-square test can be used to test categorical data. The chi-square test can not only provide whether any observed value is significantly differentbut also which category is attributable to the significant difference. The verification results are presented in an array. The equation (5) is for calculating the chi-square value:

$$\Sigma X_{i,j}^{\ 2} = \frac{(O - E)^2}{E} \quad \textbf{(5)}$$

2.4 Experiment Procedure

The experimental process of this study is based on the research discussed by [18]. The original data set of re-hospitalized diabetic patients contains incomplete and redundant information. Therefore, several attributes with extremely high missing values are not included; each patient only considers the first encounter and does not include patients who are sent to hospice or died after discharge.

There are many types of unbalanced data (Imbalanced Data). This research uses "random under-sampling" to deal with category imbalance problems for subsequent analysis. And, this study uses 10-level cross-validation on the pre-processed data to test the data to verify the accuracy of the model prediction.

3 Results

This section is divided into two main sub-sections, each of which presents the results relating to one of the research questions. The first sub-section discusses the classification results of re-hospitalization, and the second sub-section discusses the results of association rules.

3.1 Re-hospitalization Classification Results

This section will compare whether the feature variable reduction methods used as pre-processing can be improved through data mining method so as to get better classifying performance. The rough sets, Boruta and classification results of the feature variable reduction of the recursive feature reduction method will be described in the following.

3.1.1 Classification Results of Rough Set Feature Reduction

This study uses Package Rough Sets in the R software to perform rough set feature reduction on the original data after pre-processing and then conducts data exploration and analysis. The performance indicators in section 2.3 are used as the evaluation criteria. The results are shown in Table 2. The 63.57% accuracy rate of SVM is the highest, while the 59.73% accuracy rate of the C4.5 decision tree is the lowest.

Table 2. Classification results of rough set

	SVM	Naïve Bayes	C4.5	Logistic regression
Accuracy	63.57%	62.81%	59.73%	62.13%
Precision	0.639	0.621	0.635	0.625
Recall	0.623	0.656	0.457	0.611
Specificity	0.623	0.656	0.457	0.611

3.1.2 Classification Result of Boruta Feature Reduction

This study uses Package Boruta in the R software to perform Boruta feature reduction on the original data after pre-processing, and then conduct data exploration and analysis, and use the performance indicators in section 2.3 as the evaluation criteria. The results are shown in Table 3, SVM accuracy rate of 63.59% is the highest, and the accuracy of 59.67% of the C4.5 decision tree is the lowest.

Table 3. Boruta classification results

	SVM	Naïve Bayes	C4.5	Logistic regression
Accuracy	63.59%	62.22%	59.67%	62.12%
Precision	0.640	0.625	0.635	0.624
Recall	0.623	0.660	0.456	0.609
Specificity	0.623	0.660	0.456	0.609

3.1.3 Classification Results of Feature Reduction by Recursive Feature Reduction

This study uses Package Caret in the R software to perform recursive feature reduction on the original data after pre-processing, and then conduct data exploration and analysis, and use the performance indicators in section 2.3 as the evaluation criteria. The results are shown in Table 4. The 63.97% accuracy rate of SVM is the highest, and the 61.40% accuracy rate of the C4.5 decision tree is the lowest.

Table 4. Classification results of the recursive feature reduction method

	SVM	Naïve Bayes	C4.5	Logistic regression
Accuracy	63.97%	63.43%	61.40%	63.57%
Precision	0.649	0.628	0.633	0.640
Recall	0.609	0.660	0.543	0.621
Specificity	0.609	0.660	0.543	0.621

3.1.4 Comparison of Classification Results

This study compared rough set feature reduction, Boruta feature variable reduction, and recursive feature reduction method feature variable reduction in data pre-processing, and then used 4 types of data such as SVM, Naïve Bayes, C4.5 decision tree and Logistic regression. The exploration techniques are compared in a better way. The classification results are shown in Table 5. From the results of 10-level cross-validation, it can be seen that no matter which data exploration method is used, the accuracy of feature variable deletion by the recursive feature reduction method is the best, and the 63.97% accuracy of SVM is the highest.

From the perspective of data exploration methods, it can be seen that no matter which feature reduction method is used, the accuracy of SVM is the best, but the C4.5 decision tree has poor accuracy in any feature reduction method. Among them, the 59.67% accuracy rate of Boruta feature reduction is the lowest. From the above description, it is known that the recursive feature reduction method with SVM can get a better prediction accuracy.

Table 5. Comparison of classification results

Data mining method	feature reduction method	Accuracy	Precision	Recall	Specificity
SVM	Unreduction	63.15%	0.635	0.619	0.644
	Rough set	63.57%	0.639	0.623	0.623
	Boruta	63.59%	0.640	0.623	0.623
	RFE	**63.97%**	0.649	0.609	0.609
Naïve Bayes	Unreduction	63.20%	0.626	0.655	0.655
	Rough set	62.81%	0.621	0.656	0.656
	Boruta	62.22%	0.625	0.660	0.660
	RFE	**63.43%**	0.628	0.660	0.660
C4.5	Unreduction	60.62%	0.624	0.536	0.536
	Rough set	59.73%	0.635	0.457	0.457
	Boruta	59.67%	0.635	0.456	0.456
	RFE	**61.40%**	0.633	0.543	0.543
Logistic regression	Unreduction	62.03%	0.623	0.609	0.609
	Rough set	62.13%	0.625	0.611	0.611
	Boruta	62.12%	0.624	0.609	0.609
	RFE	**63.57%**	0.640	0.621	0.621

3.2 Association Rule Results

Association rule mining is an important data mining technology. This technology can mine the correlation of consumers' purchase behavior from the transaction database, find out the correlation between items, and explore interesting and relevant rules between data. This section will explore the results of Frequent Collection and Apriori in Sections 3.2.1 and 3.2.2 respectively, and find out the rules associated with the 30-day hospitalization.

3.2.1 Frequent Set Results

The purpose of exploring frequent item sets is to find out from the transactions whether the number of transactions that are purchased at the same time has a product combination that meets the user-defined threshold value, so as to combine and promote each frequent item set. The most basic method for generating frequent item sets is to calculate the frequency of occurrence (or occurrence times) of each candidate item set (Candidate Item set) of the item set point. The candidate item set is all the combination of items that may become frequent item sets [21].

In this study, the 30-day re-hospitalization data was specially selected, and the recursive feature reduction method with high classification accuracy was used to conduct the exploration from frequent one-item set to frequent five-item set to find and re- A combination of frequent item sets related to hospitalization. Figure 2 to Figure 6.

The researches of "Diabetes Control and Complications Clinical Trials" and "British Prospective Diabetes Research" found that strict blood glucose

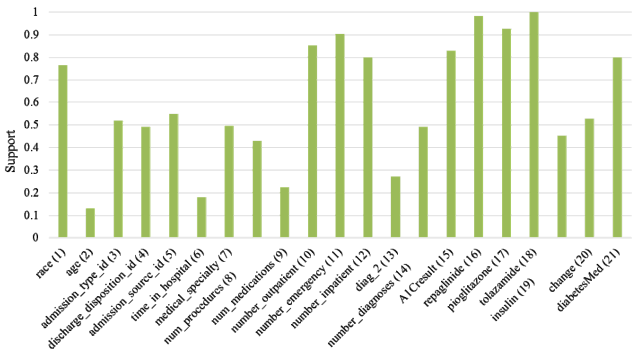


Figure 2. Frequent item set results

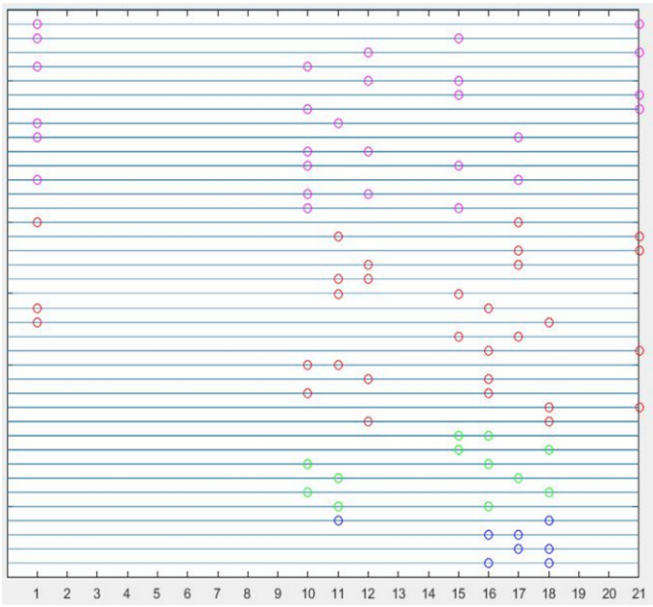


Figure 3. Frequent binomial set results

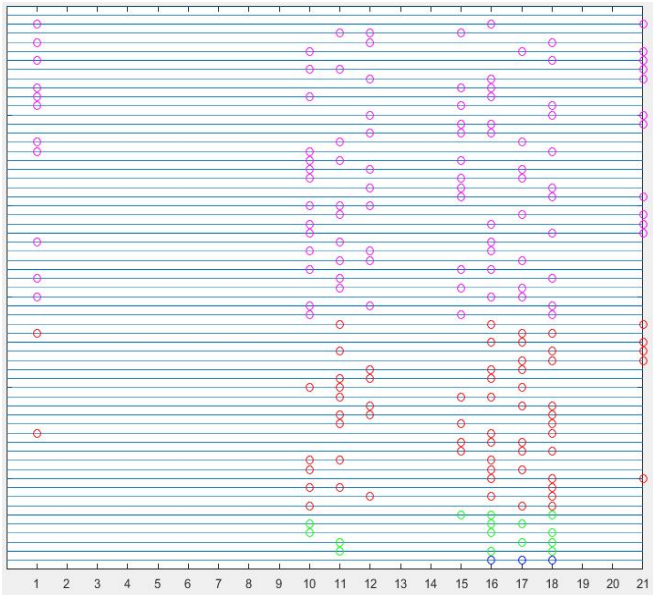


Figure 4. Frequent tri-item set results

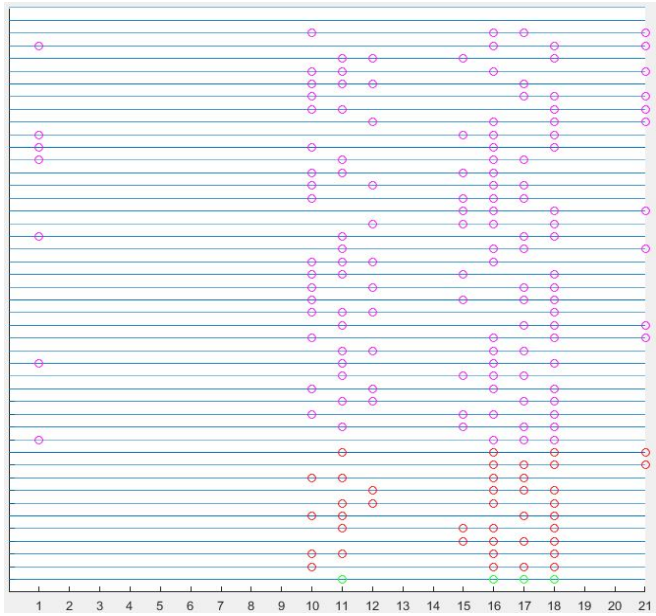


Figure 5. Frequent quartet results

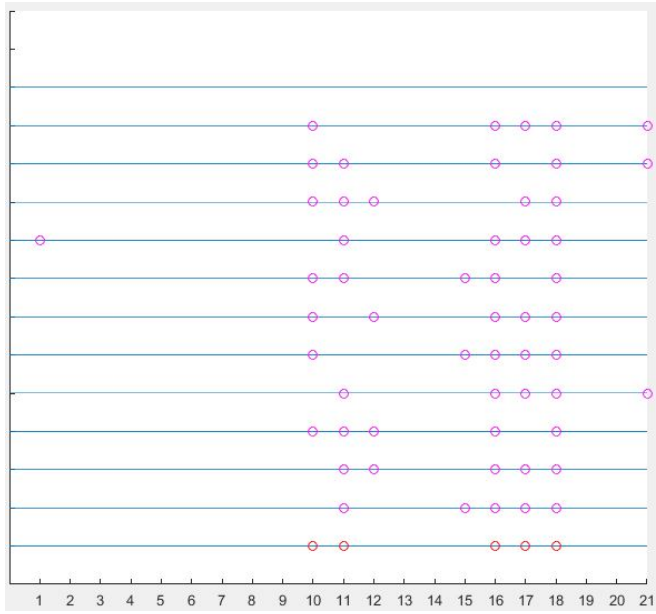


Figure 6. Frequent five itemsets results

control can effectively reduce the prevention or delay of the occurrence of diabetic complications. Therefore, this study can be obtained from the frequent collection of results. There is a high correlation between the use of blood sugar control drugs and re-hospitalization.

3.2.2 Apriori Rule Result

In this study, in order to find out the association between diabetic patients and the variables related to re-hospitalization, the recursive feature reduction method with a higher classification accuracy of re-hospitalization classification results was used to select variables.

When exploring association rules, in order to find useful associations, all association rules must meet three thresholds, namely minimum support (minimum support), minimum confidence (minimum confidence) and gain (lift). In this study, only 35 rules were selected for reliability greater than 0.9 and gain greater than 1. Sorted by the degree of trust, the best trust is 0.986. There are 4 rules, namely:

1. Number of outpatient visits = 0 & number of emergency visits = 0 &repaglinide = No & number of hospitalizations = 0 → will be hospitalized again in 30 days
2. The number of outpatient visits = 0 & the number of emergency visits = 0 &repaglinide = No & Tolazamide = No & the number of hospitalizations = 0 → will be hospitalized again in 30 days
3. Number of outpatient visits = 0 & Number of hospitalizations = 0 & Pioglitazone = No & Tolazamide = No & Number of emergency visits = 0 → Will be hospitalized again in 30 days
4. Number of outpatient visits = 0 & number of emergency visits = 0 & tolazamide = No & number of hospitalizations = 0 → will be hospitalized again in 30 days

3.3 CART Decision Tree Analysis Results

This research uses Package rpart in R software to analyze the CART decision tree. This study tried to set a few more complexity parameter values to view the nodes of the CART decision tree, and finally established two CART decision trees whose complexity parameter values were set to 0.01 and 0.003 to assess the risk of re-hospitalization. And according to the International Disease Classification Code Reference Table of the Ministry of Health and Welfare, the values in the auxiliary diagnosis variables are divided into categories, and a total of 16 categories are divided to successfully analyze the CART decision tree.

As shown in Figure 7, the value of the complexity parameter is the CART decision tree set to 0.01, and this study divided the five results of the CART decision tree into high, medium, and low risk groups for re-hospitalization, high risk group 1, group 1, Medium risk group 2 group and low risk group 2 group. Table 6 lists the chances of re-hospitalization for each risk group. Among patients in the high-risk group, if the value of the treatment variable after discharge is between 11 and 22.5, as high as 85.15%, they are likely to be re-hospitalized. In this study, to check whether there are significant differences between groups, the risk measurement Odds ratio and chi-square test are used to compare risk groups. The Odds ratio between the highest risk group and the lowest risk group is 21.25. Both have significant differences, as shown in Table 7.

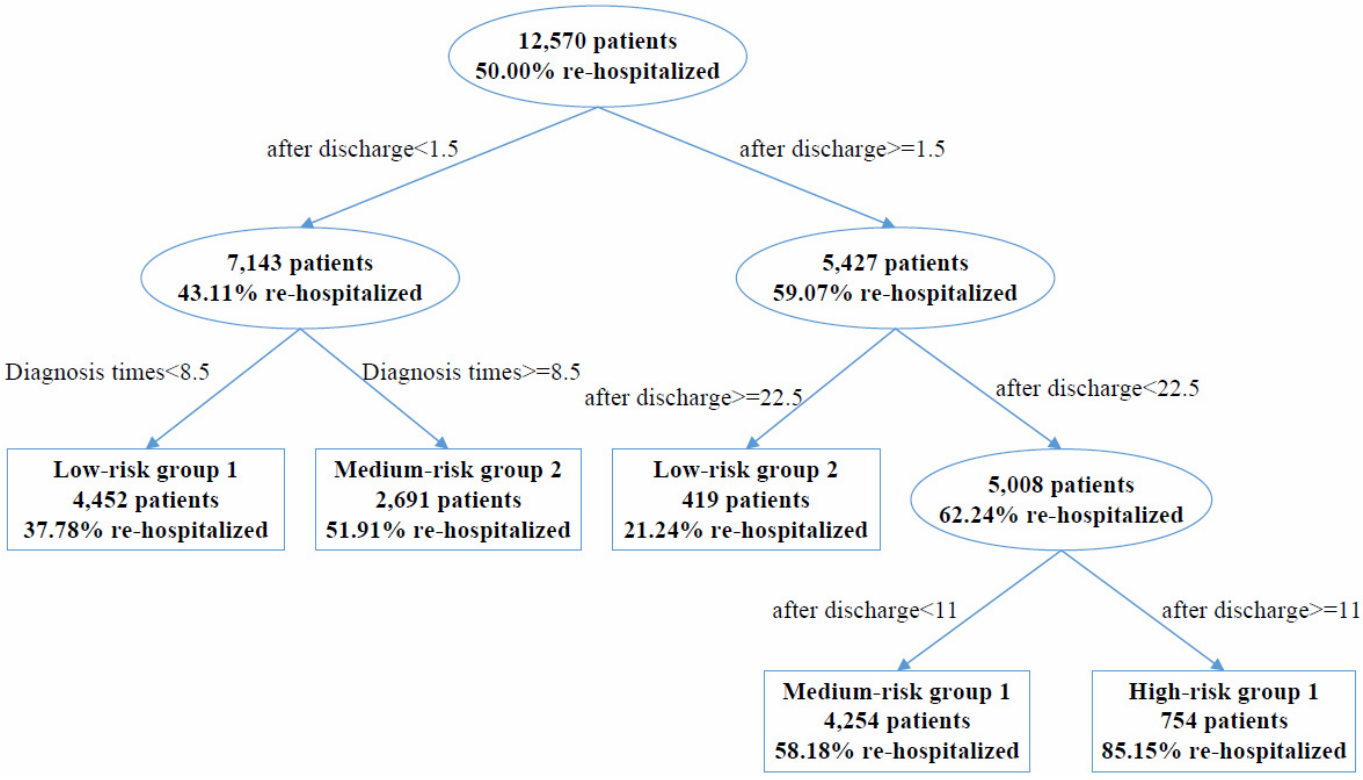


Figure 7. Re-hospitalization risk stratification model based on CART decision tree_parameter 0.01

Table 6. Re-hospitalization rate of risk group

Risk Group	Re-hospitalized population/population of the group
High 1	642/754 (85.15%)
Medium 1	2475/4254 (58.18%)
Medium 2	1397/2691 (51.91%)
Low 1	1682/4452 (37.78%)
Low 2	89/419 (21.24%)

Table 7. Comparison of Odds ratio among risk groups

Risk group analysis	Odds ratio
High 1 vs	
Low 2	21.25
Low 1	9.44
Medium 2	5.31
Medium 1	4.12
Medium 1 vs	
Low 2	5.16
Low 1	2.29
Medium 2	1.29
Medium 2 vs	
Low 2	4.00
Low 1	1.78
Low 1 vs	
Low 2	2.25

*$p < .001$.

As shown in Figure 8, the value of the complexity parameter is the CART decision tree set to 0.003. In this study, the 11 results of the CART decision tree are divided into high, medium, and low risk groups for re-hospitalization, high risk group 2, medium Risk group 6 groups and low risk group 3 groups. Table 8 lists the chances of re-hospitalization for each risk group. Among patients in high-risk group 1, the value of the treatment variable after discharge is greater than 26.5 and up to 94.12% may be re-hospitalized. In the high-risk group 2 patients, the value of the treatment variable after discharge is between 11 and 22.5, as high as 85.15% are likely to be re-hospitalized. In order to check whether there are significant differences between groups, the risk measurement Odds ratio and Chi-square test are used to compare risk groups. The Odds ratio between the highest risk group and the lowest risk group is 92.07. Most of the groups have significant differences, as shown in Table 9.

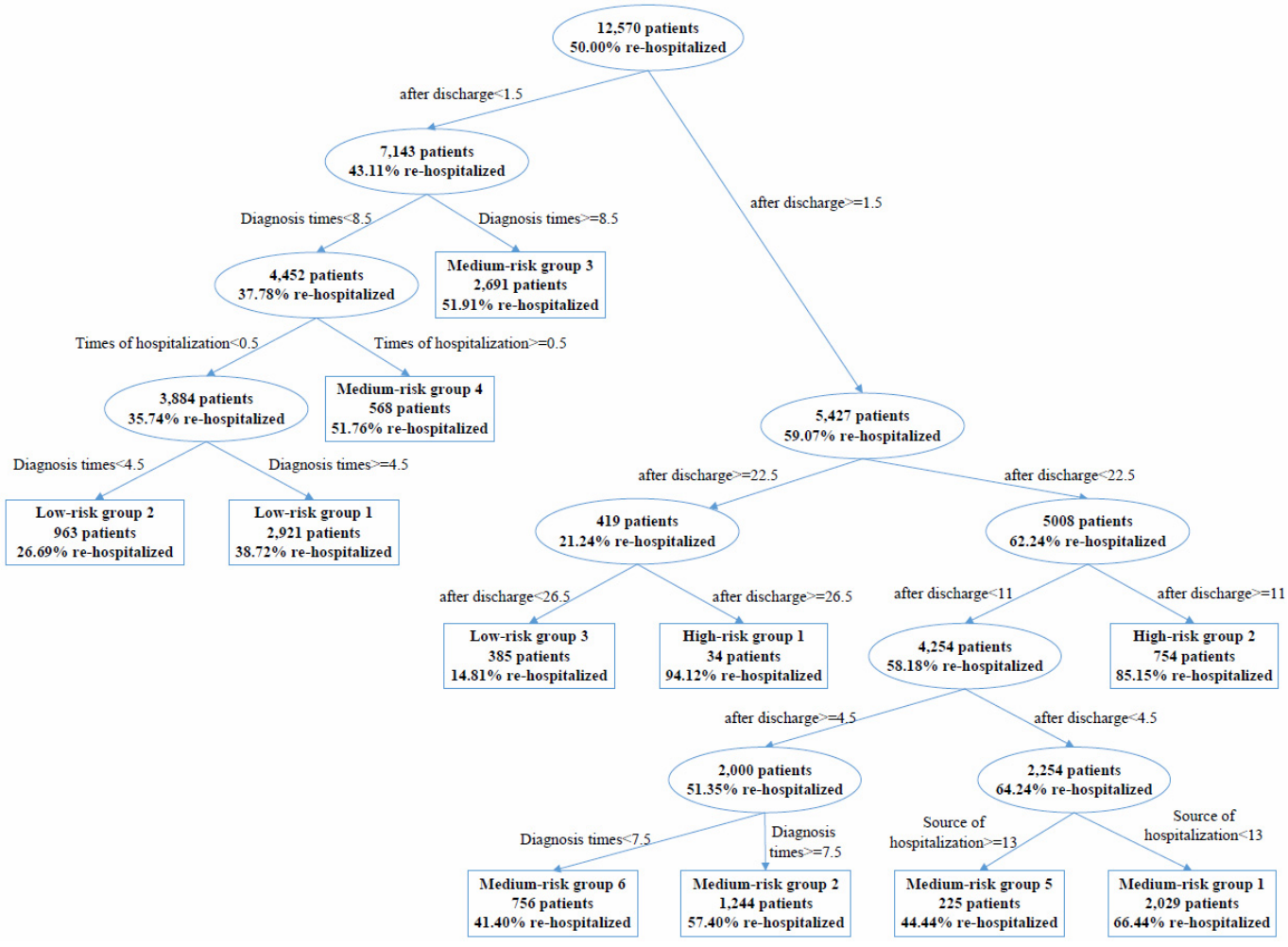


Figure 8. Re-hospitalization risk stratification model based on CART decision tree_parameter 0.003

Table 8. Re-hospitalization rate of risk group

Risk Group	Re-hospitalized population/population of the group	Risk Group	Re-hospitalized population/population of the group
High 1	32/34 (94.12%)	Medium 5	100/225 (44.44%)
High 2	642/754 (85.15%)	Medium 6	313/756 (41.40%)
Medium 1	1348/2029 (66.44%)	Low 1	1131/2921 (38.72%)
Medium 2	714/1244 (57.40%)	Low 2	257/963 (26.69%)
Medium 3	1397/2691 (51.91%)	Low 3	57/385 (14.81%)
Medium 4	294/568 (51.76%)		

3.4 Comparison of Experimental Results

In this section, this research has compiled and compared with other scholars' research, as shown in Table 10. The results of re-hospitalization classification are measured by the performance indicators in section 2.3 to measure the accuracy of prediction. The SVM method of this study has the highest accuracy of 63.97%. This study uses three feature variable reduction methods to remove variables that do not affect the classification results. The experimental results show that the recursive feature removal method feature variable reduction and support vector machine can get better prediction accuracy.

Table 9. Comparison of Odds ratio among risk groups

Risk group analysis	Odds ratio	*p* value
High 1 vs		
Low 3	92.07	<.001*
Low 2	43.95	<.001*
Low 1	25.32	<.001*
Medium 6	22.65	<.001*
Medium 5	20.00	<.001*
Medium 4	14.91	<.001*
Medium 3	14.82	<.001*
Medium 2	11.88	<.001*
Medium 1	8.08	<.001*
High 2	2.79	.210
High 2 vs		
Low 3	32.99	<.001*
Low 2	15.75	<.001*
Low 1	9.07	<.001*
Medium 6	8.11	<.001*
Medium 5	7.17	<.001*
Medium 4	5.34	<.001*
Medium 3	5.31	<.001*
Medium 2	4.26	<.001*
Medium 1	2.90	<.001*
Medium 1 vs		
Low 3	11.39	<.001*
Low 2	5.44	<.001*
Low 1	3.13	<.001*
Medium 6	2.80	<.001*
Medium 5	2.47	<.001*
Medium 4	1.85	<.001*
Medium 3	1.83	<.001*
Medium 2	1.47	<.001*

Table 9. Comparison of Odds ratio among risk groups (continue)

Risk group analysis	Odds ratio	*p* value
Medium 2 vs		
Low 3	7.75	<.001*
Low 2	3.70	<.001*
Low 1	2.13	<.001*
Medium 6	1.91	<.001*
Medium 5	1.68	<.001*
Medium 4	1.26	.028*
Medium 3	1.25	.001*
Medium 3 vs		
Low 3	6.21	<.001*
Low 2	2.97	<.001*
Low 1	1.71	<.001*
Medium 6	1.53	<.001*
Medium 5	1.35	<.001*
Medium 4	1.01	.963
Medium 4 vs		
Low 3	6.17	<.001*
Low 2	2.95	<.001*
Low 1	1.70	<.001*
Medium 6	1.52	<.001*
Medium 5	1.34	.070
Medium 5 vs		
Low 3	4.60	<.001*
Low 2	2.20	<.001*
Low 1	1.27	.103
Medium 6	1.13	.442
Medium 6 vs		
Low 3	4.07	<.001*
Low 2	1.94	<.001*
Low 1	1.12	.181
Low 1 vs		
Low 3	3.64	<.001*
Low 2	1.74	<.001*
Low 2 vs		
Low 3	2.10	<.001*

In the results of association rules, this study uses frequent sets and Apriori algorithm to find association rules. The results show that the use of blood sugar control drugs is highly correlated with re-hospitalization. Therefore, if patients are not taking blood sugar lowering drugs, there is a great possibility that you will be hospitalized again. In the experimental result comparison table in Table 10, this study analyzes the risk of re-hospitalization, that is, the CART decision tree analysis result in Section 3.3, where the value of the treatment variable after discharge is greater than 26.5 and up to 94.12% may be re-hospitalized, which is a high risk The group of patients must pay great attention to their physical condition.

4 Conclusion

With the increasing cost of diabetes to the health care system, re-hospitalization is recommended as a goal to reduce health care costs. This paper use three feature reduction methods to compare and use four

Table 10. Comparison of experimental results

author	research method	Pretreatment method	Association rules	Accuracy
Strack et al. (2014)	Logistic regression	• Ignore attributes with high missing values • Only the patients with the first encounter are adopted, and the hospice care and death are deleted	The significant relationship between readmission and glycated blood red and white depends on the primary diagnosis	N/A
Hephzibah et al. (2015)	SVM	• The PROC SQL analysis removes the attributes with missing values • SAS code reduction properties	The key factors of re-hospitalization are the initial diagnosis, hospitalization mode, etc.	61.6%
Bhuvan et al. (2016)	Naïve Bayes	• The ICD9 codes of similar diagnosis are divided into 10 groups • Ignore attributes with high missing values	Patients diagnosed with cellulitis / abscess and without special complications have a very low risk of re-hospitalization	0.214 (Recall)
	Bayesian network			0.208 (Recall)
	Random Forest			0.242 (Recall)
	Adaboost			0.167 (Recall)
	ANN			0.233 (Recall)
This study	SVM	• Ignore attributes with high missing values • Class imbalance data processing • Feature reduction method eliminates more attributes	1. If you don't take hypoglycemic drugs, you may be hospitalized again 2. The value of disposition variable after discharge is greater than 26.5, and there is a high risk of re-hospitalization as high as 94.12%	63.97%
	Naïve Bayes			63.43%
	C4.5			61.40%
	Logistic regression			63.57%
	CART			N/A

data mining methods to select the best predictive model. At the same time, we use Apriori algorithm and CART decision tree to analyze the results of re-hospitalization classification, association rule results, and re-hospitalization risk.

The experimental results showed that the recursive feature reduction method combined with SVM can get a better prediction accuracy. In the association rule result, the frequent set and the Apriori algorithm are used to find the association rule, and the result is whether the drug to control blood sugar is used or not, re-hospitalization is highly correlated. The experimental results of this study can grasp the possibility of patients being re-hospitalized within 30 days, and remind patients of what needs to be paid more attention to in order to reduce the possibility of increased mortality and the burden of medical costs.

References

[1] Ministry of Health and Welfare, https://www.mohw.gov.tw/cp-16-54482-1.html

[2] The International Diabetes Federation (IDF), *IDF Diabetes Atlas*, 8th ed., 2017, p. 41.

[3] I. Sayago-Silva, F. García-López, J. Segovia-Cubero, Epidemiología de la insuficiencia cardiaca en España en los últimos 20 años, *Revista Española de Cardiología*, Vol. 66, No. 8, pp. 649-656, August, 2013.

[4] J. Wu, W. Guo, Y. Tang, A. Han, B. Yang, D. Zhang, B. Zhang, A Study of TCM Master Yan Zhenghua's medication Rule in Prescriptions for Digestive System Diseases Based on Apriori and Complex System Entropy Cluster, *Journal of Traditional Chinese Medical Sciences*, Vol. 2, No. 4, pp. 241-247, October, 2015.

[5] M. J. Swain, H. Kharrazi, Feasibility of 30-day Hospital Readmission Prediction Modeling Based on Health Information Exchange Data, *International Journal of MEdical Informatics*, Vol. 84, No. 12, pp. 1048-1056, December, 2015.

[6] S. Howell, M. Coory, J. Martin, S. Duckett, Using Routine Inpatient Data to Identify Patients at Risk of Hospital Readmission, *BMC Health Services Research*, Vol. 9, No. 1, pp. 1-9, June, 2009.

[7] T. Fraze, H. J. Jiang, J. Burgess, Hospital Stays for Patients with Diabetes, 2008, *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*, pp. 1-11, August, 2010.

[8] American Diabetes Association, Economic Costs of Diabetes in the US in 2012, *Diabetes care*, Vol. 36, No. 4, pp. 1033-1046, April, 2013.

[9] H. J. Jiang, D. Stryer, B. Friedman, R. Andrews, Multiple Hospitalizations for Patients with Diabetes, *Diabetes Care*, Vol. 26, No. 5, pp. 1421-1426, May, 2003.

[10] D. J. Rubin, Hospital Readmission of Patients with Diabetes, *Current Diabetes Reports*, Vol. 15, No. 4, pp. 1-9, April, 2015.

[11] J. Futoma, J. Morris, J. Lucas, A Comparison of Models for Predicting Early Hospital Readmissions, *Journal of Biomedical Informatics*, Vol. 56, pp. 229-238, August, 2015.

[12] R. J. Lagoe, C. M. Noetscher, M. P. Murphy, Hospital Readmission: Predicting the Risk, *Journal of Nursing Care Quality*, Vol. 15, No. 4, pp. 69-83, July, 2001.

[13] J. Billings, J. Dixon, T. Mijanovich, D. Wennberg, Case Finding for Patients at Risk of Readmission to Hospital:

Development of Algorithm to Identify High Risk Patients, *MJ*, Vol. 333, No. 7563, Article number 327, August, 2006.

[14] P. T. Donnan, D. W. Dorward, B. Mutch, A. D. Morris, Development and Validation of a Model for Predicting Emergency Admissions over the Next Year (PEONY): A UK Historical Cohort Study, *Archives of Internal Medicine*, Vol. 168, No. 13, pp. 1416-1422, July, 2008.

[15] C. van Walraven, J. Wong, S. Hawken, A. J. Forster, Comparing Methods to Calculate Hospital-specific Rates of Early Death or Urgent Readmission, *Canadian Medical Association Journal*, Vol. 184, No. 15, pp. E810-E817, October, 2012.

[16] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, S. Kripalani, Risk Prediction Models for Hospital Readmission: A Systematic Review, *Jama*, Vol. 306, No. 15, pp. 1688-1698, October, 2011.

[17] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, S. Poranki, Predictive Modeling of Hospital Readmissions Using Metaheuristics and Data Mining, *Expert Systems with Applications*, Vol. 42, No. 20, pp. 7110-7120, November, 2015.

[18] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, J. N. Clore, Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, *BioMed Research International*, Vol. 2014, Article ID 781670, April, 2014.

[19] M. Agaoglu, Predicting Instructor Performance Using Data Mining Techniques in Higher Education, *IEEE Access*, Vol. 4, pp. 2379-2387, May, 2016.

[20] N. Seliya, T. M. Khoshgoftaar, J. Van Hulse, A Study on the Relationships of Classifier Performance Metrics, *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, Newark, NJ, USA, 2009, pp. 59-66.

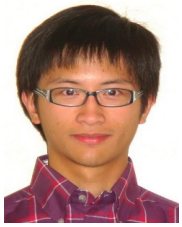
[21] J.-W. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan-Kaufmann Academic Press, 2001.

Biographies

Dong-Her Shih received his Ph.D. degree in Electrical engineering from National Cheng Kung University, Taiwan, in 1986. He is a senior professor in Department of Information Management, National Yunlin University of Science and Technology, Douliu, Yunlin, Taiwan. Dr. Shih is the Associate Editor of the *International Journal of Mobile communication* and published over 70 journal articles. His current researches include data mining, information security, Block Chain and Big data.

Feng-Chuan Huang received his M.S. degree from Department of Information Management, Da-Yeh University, Taiwan, in 2007. She is currently a Ph.D. student in Department of Information Management, National Yunlin University of Science and Technology, Taiwan. His current researches include Data mining and Big data analysis.

Cai-Ling Weng received his M.S. degree from Department of Information Management, National Yunlin University of Science and Technology, Douliu, Yunlin, Taiwan, in 2017. His major research includes Computer security and Data mining.



Po-Yuan Shih received his M.S. degree from Department of Finance, National Formosa University, Taiwan, in 2015. He is currently a Ph.D. student in Department of Finance, National Yunlin University of Science and Technology, Taiwan. His major research includes Data mining, FinTech and Marketing.

David C. Yen is currently a Professor at Texas Southern University. Professor Yen is active in research and has published books and articles which have appeared in ACM Transaction of MIS, Decision Support Systems, Information & Management, Decision Sciences, International Journal of Electronic Commerce, ACM SIG Data Base, Information Sciences, Communications of the ACM, Government Information Quarterly, IEEE IT Professionals, Information Society, Omega, International Journal of Organizational Computing and Electronic Commerce, and Communications of AIS among others. Professor Yen's research interests include e-government, mobile commerce, medical information systems, and IT auditing/governance.