مجلة جامعة بني وليد للعلوم الإنسانية والتطبيقية
تصدر عن جامعة بني وليد – ليبيا
Website: https://jhas-bwu.com/index.php/bwjhas/index
المجلد التاسع، العدد الأول 2024

المصارف وتعدين البيانات (باستخدام أداتين مختلفتين)

عبدالسلام رحيل اكرومة
قسم الحاسوب، كلية العلوم، جامعة بني وليد، ليبيا
abdalsalamekroma@bwu.edu.ly

Data Mining and Banking (Using two different tools)

Abdulssalam Jomah Akroma
Computer Department, Faculty of Science, Bani Waleed University, Libya

تاريخ الاستلام:2024-02-09 تاريخ القبول: 2024-02-28 تاريخ النشر: 2024-03-09

الملخص:

التحدي الأكبر الذي يواجهه المصرفيون هذه الأيام هو كيفية تحسين أساليب جمع البيانات للحصول على المعلومات ذات الصلة؛ علاوة على ذلك، كيفية تنظيم وتحليل البيانات المجمعة بشكل فعال. إن تبني الأساليب الصحيحة سيساعد المصرفيين في نهاية المطاف على اتخاذ القرار المناسب لتحقيق أهدافهم المرجوة والحصول على إجابات لقضايا العمل الرئيسية، بالإضافة إلى أنها ستساعدهم على التنبؤ بسلوك العملاء الجدد.

والهدف من المشروع هو تحليل مجموعة بيانات بنك ومعرفة عوامل المدخلات التي كان لها أكبر تأثير على النتيجة سواء كانت مرغوبة أم غير مرغوب فيها.

الكلمات الدالة: التصنيف، أداة **RapidMiner**، استخراج البيانات، إضافة **SQL** للإكسيل ، تعدين البيانات.

Abstract

The biggest challenge that bankers face these days is how to improve the data collection methods to get relevant information; moreover, how to organize and analyze the collected data effectively. Adopting the right methods will ultimately help bankers to take the appropriate decision to achieve their desired goals and to get the answers of the main business issues, besides that it will help them to predict new customer's behavior. The objective of this project is to analyze a bank dataset and find out what input factors had a greatest impact on the outcome whether it was desirable or not desirable.

Keywords: Classification, RapidMiner tool, data extraction, adding SQL to Excel, data mining.

Introduction:

One of many areas in which techniques of Business intelligence (BI) and Data Mining (DM) can bring improvement is targeted marketing campaign that has been heavily invested recently by marketing managers and in which Data Mining techniques based on the naive bayes classifier

has been implemented. This implementation yields better analysis of client's behavior and success of contact; explaining whether the client registers the deposit or not. Subsequently, this implementation leads marketing managers to discover main attributes that encourage clients to register deposit, thereby allowing efficient management. In other words, such DM project assists marketing managers to make more efficient use of resource and time in their disposal to select customers who have more potential to register deposit. [1]

:Background

Due to globalization and economic competitiveness that global open market brings, there has been a need for more efficient marketing method. Using mass campaigns that address general public has been regarded as gradually inefficient due to public's irresponsiveness and lack of positive response towards such campaigns. [2] Instead, marketing managers have been turning to usage of direct marketing that addresses and chooses customers that are more interested or will be more responsive to specific product or service. For example, many European banks which sought to increase their financial asset to secure themselves from financial crisis have followed conventional economic knowledge and, therefore, provided long-term, good interest rate deposit applications through direct marketing campaigns. However, using direct marketing campaigns does not guarantee maximum efficiency. Rather, regarding cost-benefit and time-benefit efficiency, it was noted that in order to improve efficiency, less contacts should be made while a number of clients registering the deposit should either increase or be unchanged. To increase efficiency, several DM algorithms have been used, including Decision Tree (DT), Naïve Bayes (NB), and association rules have been applied. Their implementation to the area of marketing campaigns allows construction of a prediction model that categorizes an element of data into a predefined class, thereby classifying different marketing contacts. Then, are the classifiers used in these techniques accurate? Although many assessment tools can be used, the most favored one for marketing campaigns is RapidMiner tool.

1.1 Data Mining:

Data Mining (DM) corresponds to the definition and idea of "mining" itself: briefly, DM is a logical process of "mining" or obtaining practical and valuable and relevant information from large set of data. It can yield not only well-informed decision based on mined information and knowledge but also predictive analyses on future patterns and trends. The fact that DM is not a simple computerized and engineered method but a combinatory process of math and statistics shows the profundity of its resulting information and predictive ability. [1]

1.2 Banks and Data mining

As banks accumulate enormous volume of data on their assets, customers and other financial elements, there is a need to extract valuable information from such data to make efficient decisions. For banks and other financial institutions for that matter, profit maximization is the primary goal. How do banks accomplish such objective? Indeed, answer vary greatly: constant marketing to expand their customer bases, offering attractive investment scheme, avoiding possible default, and predicting prospective current in financial market to prepare for deflation, inflation and other regressions. Then, another question begs whether DM can address and facilitate such numerous tasks. [1]Unfortunately, DM cannot mitigate economic and market uncertainties completely or maximize profits to an unprecedented level. However, as mentioned, DM allows more efficient use of capital, financial and human resources in banks' disposal. For example, a valuable pattern and knowledge that are extracted from a data may inform decision makers to design an appropriate loan plan with attractive interest rate that corresponds to prevalent economic current, thereby garnering more customers. Subsequently, predictive analysis may inform financial institutions of possible rate or risk of defaulting, thereby mitigating risk of financial loss. In addition, predictive analysis may advise banks when to free up capital for additional investment and provide insight on future stock markets. All these not only maximize profit and mitigate possible risk but also encourage efficient arrangement of investment scheme to gather more investors [3].

1.3 What Kind of Data Does a Bank Need to Engage in Data mining?

It has been mentioned that a bank accumulates a large amount of data. These data range from financial statistics to general information on customers. Since using every data available for DM is unpractical, let alone impossible, a type of data a bank has to use depends on the objective of the bank that DM will be addressing. For example, in order to avoid default and design better loan plan, what a bank needs are data containing information of defaulters, credit, precedent cases of default and *etc*. If a bank seeks to obtain predictive analysis on future stock performance, it needs data on past stock performance, profiles of stock holders, historic and current trends, and other stock-related data on which data mining can benefit. Outside sources, such as questionnaires and polls, can also be useful resource for implementation of DM. [1]

2. Problem Definition

Banks have a huge amount of data belongs to their clients. The biggest challenge is how to organize this kind of data in a way that can provide bankers with a reliable basis for effective decisions-making and also with predictions of the new customer behavior. Moreover, having

better understanding of the Data is very important to do so I need to understand the rules of the data.

2.1 Type of the Dataset

The used Dataset had been generated through multiple targeted campaigns by calling the clients once, twice and sometimes multiple phone calls were made to promote and reinforce the campaign. As a result; a 45212 records were registered and a dataset with useful information was generated which includes some information about the client profile like age, job, marital status and education. In addition, it includes his/her financial situation in terms of credit, account balance, and housing, also it includes number of calls that were made per customer, his/her previous response and the duration of each call per sec.

3. Proposed Solution

With this kind of data, one of the best ways to classify and predict its outcome is by using Naive Bayes classifier. However, a profound understanding of the data is required to get the more accurate results and this can be done by using the Decision tree classifier. Based on the above two facts I had used the previous mentioned algorithms. On the other hand, the association rule was also applied to find out what kind of rules are existed and could be useful ones in the campaign.

In order to enrich our results and also to enhance our technical understanding capabilities I decided to use two different tools or applications. The first tool was Rapid Miner which is one of the most well-known tools that has been used by many miners and analysts, besides that it's an open source one, while the second application was the Microsoft SQL 2018 Analysis Service with the built in data mining add-ins for MS Excel 2018.

3.1 Naive Bayes classifier

Naive Bayes is considered as one of the most efficient learning algorithms for machine learning and data mining. The naive Bayes classifier uses Bayes' Theorem, a principle that using a historical data to find out the probability value of class attribute by counting the frequency of values and combinations [4].Moreover, Naive Bayes algorithm is based on conditional probabilities and also is called conditional independence, with a naive Bayes all attributes are independent given the value of the class variable. For example, a product could be considered to be a computer if it comes from Toshiba company, cost 1000$, and includes windows 7 Operating system with it. Even if these features depend on each other, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this product is a computer. [4]

3.2 Decision tree classifier

Decision tree contains every possible outcome that appears in arrange of data. There are some advantages of using Decision tree than others methods of classifier

construction one of them is that decision trees are much less computationally intensive, relevant attributes are selected automatically. The main object of using it to predict the value of a target element based on several

input elements [2]. Figure-1 shows an example of the result of Decision tree

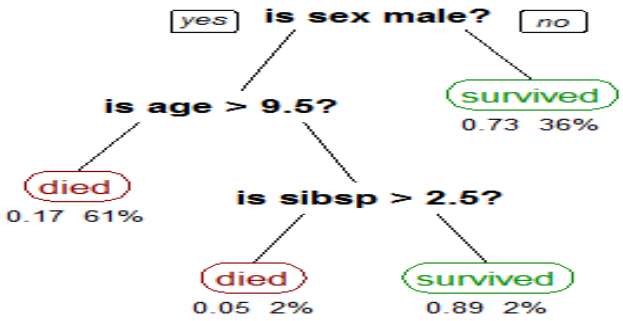


Figure -1 Tree Titanic survivors http://en.wikipedia.org/wiki/Decision_tree_learning

3.3 **Association Rule**

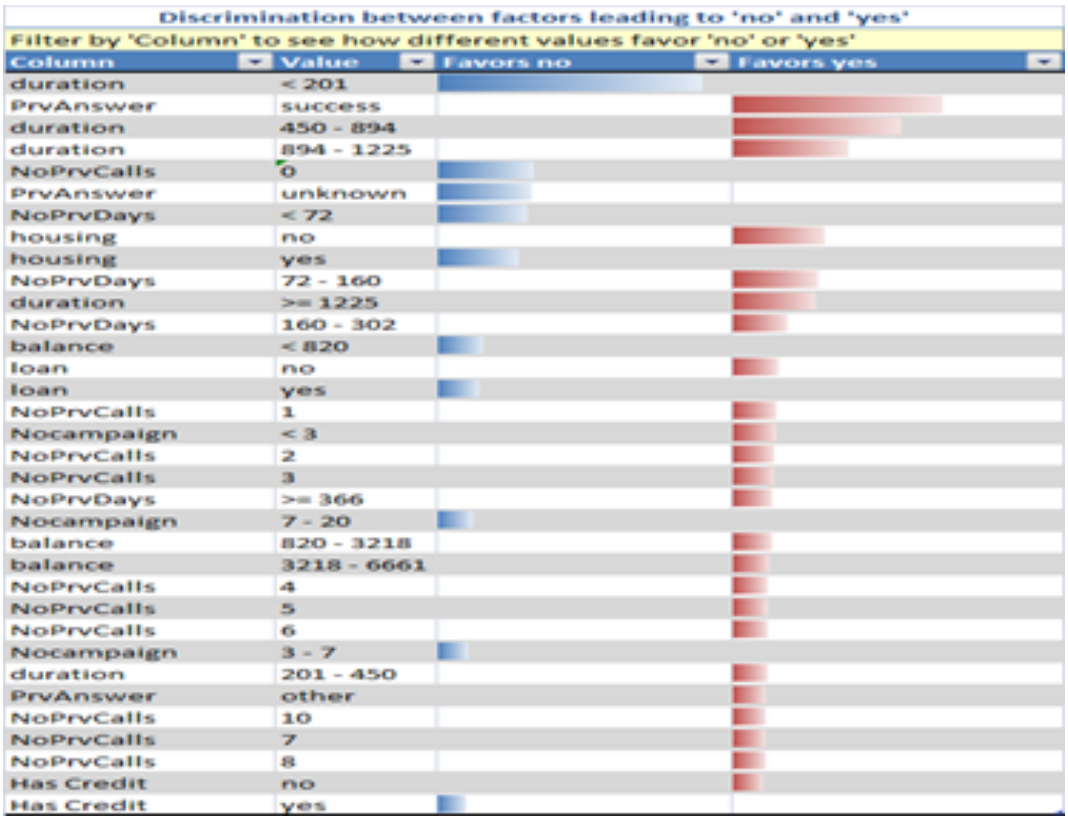
Association Rule is one of the Data Mining techniques that scan the data set to find interesting associations and correlations among its attributes value sets. This algorithm made data exploration much easier as instead of writing hundreds or thousands of queries to explore all the possible combination among the dataset was need only to run this algorithm [5]. The outcome of this rule is beneficial as it can help managers or decision makers to make effective marketing plans that would be very imperative in terms of marketing and business objectives. There are many algorithms that implement association rule, and in this project was used in particular FP-growth (frequent pattern growth Algorithm). In association rules was need to satisfy two predefined factors which are the minimum support and the minimum confidence.

4. Related work

4.1 SQL 2018 add-ins for MS Excel:

The data in the mining structure had been split into a training set and a testing set in where the data mining engine would use the training set to train the mining models, and the testing set to test the accuracy of the models. In this application was used one more tool which is the key influencer which was just to take a global idea about the key factors which would influence the outcome, although was know very well that this tool is not efficient enough as it doesn't look for

the interaction between different influencers as decision tree does. However, it could provide us with a preliminary image by ranking the relative impacts of the factors of the input which is depicted below that the duration of the call and the previous answer were one of the greatest influencers on the client's outcome.



Discrimination between factors leading to 'no' and 'yes'

Filter by 'Column' to see how different values favor 'no' or 'yes'

Column	Value	Favors no	Favors yes
duration	< 201		
PrvAnswer	success		
duration	450 - 894		
duration	894 - 1225		
NoPrvCalls	0		
PrvAnswer	unknown		
NoPrvDays	< 72		
housing	no		
housing	yes		
NoPrvDays	72 - 160		
duration	>= 1225		
NoPrvDays	160 - 302		
balance	< 820		
loan	no		
loan	yes		
NoPrvCalls	1		
Nocampaign	< 3		
NoPrvCalls	2		
NoPrvCalls	3		
NoPrvDays	>= 366		
Nocampaign	7 - 20		
balance	820 - 3218		
balance	3218 - 6661		
NoPrvCalls	4		
NoPrvCalls	5		
NoPrvCalls	6		
Nocampaign	3 - 7		
duration	201 - 450		
PrvAnswer	other		
NoPrvCalls	10		
NoPrvCalls	7		
NoPrvCalls	8		
Has Credit	no		
Has Credit	yes		

Figure −2 Key influencers output−SQL 2018

The second algorithm was the decision tree which also confirmed that the previous mentioned factors were the greatest influencers, while this method is considered as more accurate and scalable as it tests the interaction between different influencers and partitions the data based on the attribute that has the most influence.

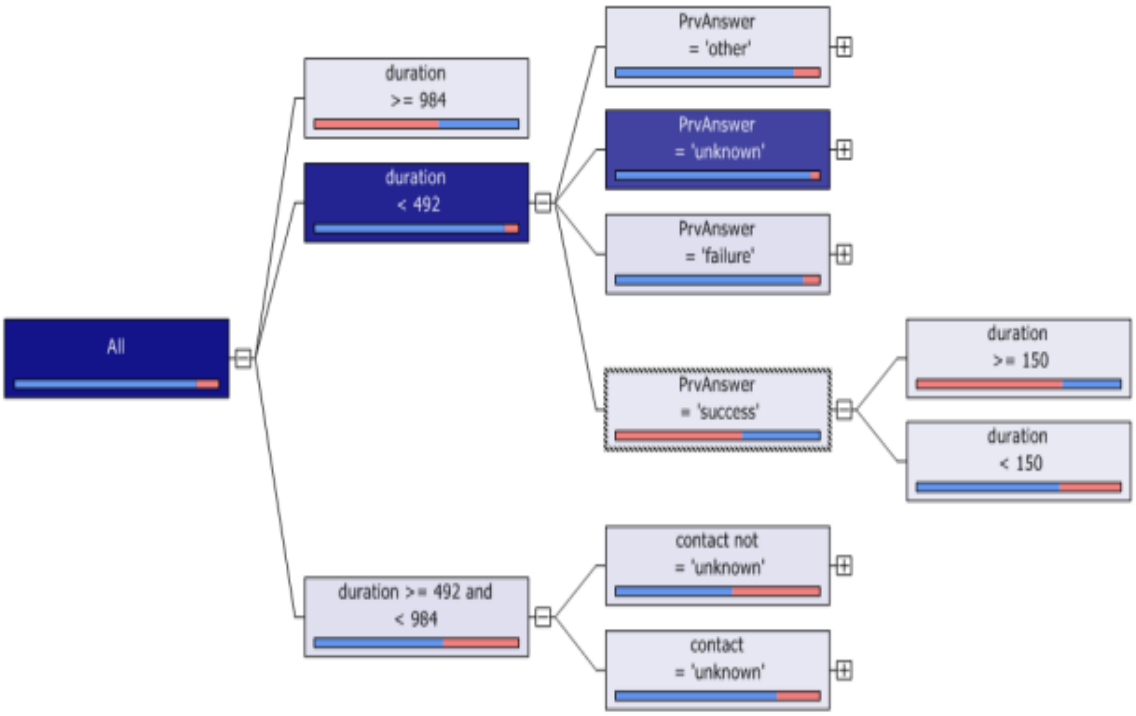


Figure-3 Decision tree output-SQL 2018

A query tool had also used from the Model usage by creating a new data without the outcome that was already included with the original dataset. By running this tool got two outputs the prediction and the probability and the objective behind this test was just to evaluate our work and the running tasks by comparing the outputs that I got to the actual outcome and also to the decision tree results. for example on duration < 492sec, previous answer: unknown, contact : unknown and month=may , I got 99.84 % from prediction query and from the decision tree as well .Excel sheet and two snapshots will be provided in the accompanied CDROM.

In SQL and data mining Add-ins for Excel Naive Bayes couldn't be used because some attributes were not supported by its content type so I had used Microsoft Neural network instead.

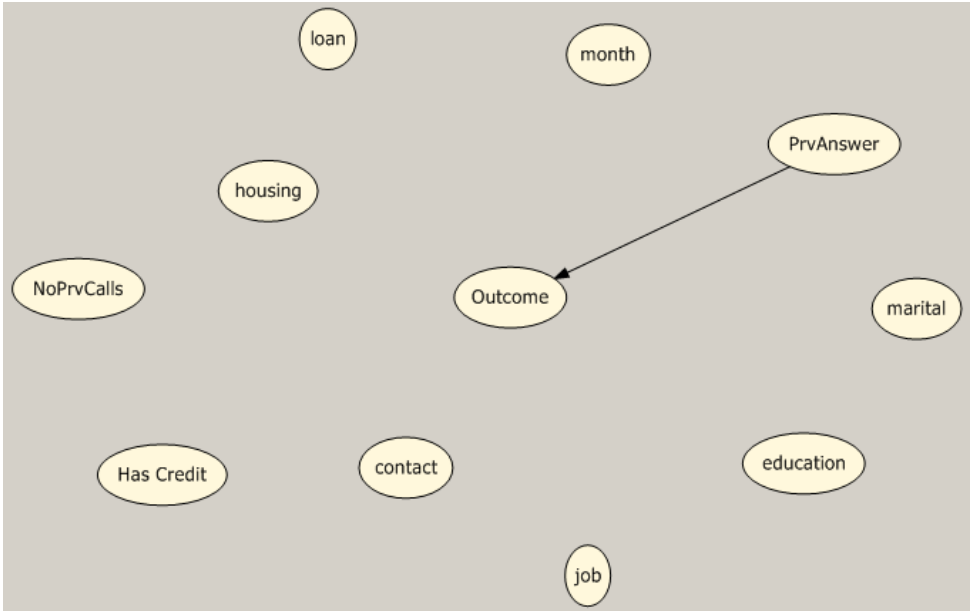


Figure-4 shows the strongest relationship between nodes

4.2 Rapid Miner

In this section, the explain the experiments performed and the accuracy of the results obtained over a dataset of the bank.

4.2.1 Accuracy and predication

The Data set was divided into parts. First part to be used as a training set and the second part to be used as a testing set. In the Data set was assumed that this dataset belongs to new clients and was do not know whether they will say yes to the campaign or they will reject it. The objective is to find which clients is more likely willing to accept this campaign based on training data. Figure 5

shows the Performance and accuracy prediction of the Dataset

accuracy: 87.64% +/- 0.65% (mikro: 87.64%)

	true no	true yes	class precision
pred. no	36840	2505	93.63%
pred. yes	3082	2784	47.46%
class recall	92.28%	52.64%	

Figure-5 Accuracy and Prediction-Rapid Miner

4.2.2 Association Rule

The association rule was also applied to find out any correlation between items that I can use to get the maximum outcome from the campaign. The Association algorithms usually deal with Non-numerical data and since I have the age, balance, etc. I used a discretize filter to convert this data to Non-numerical. For example I divide the ages into many categories such as young, youth and middle-aged, etc. Figure 6 shows the results that were obtained.

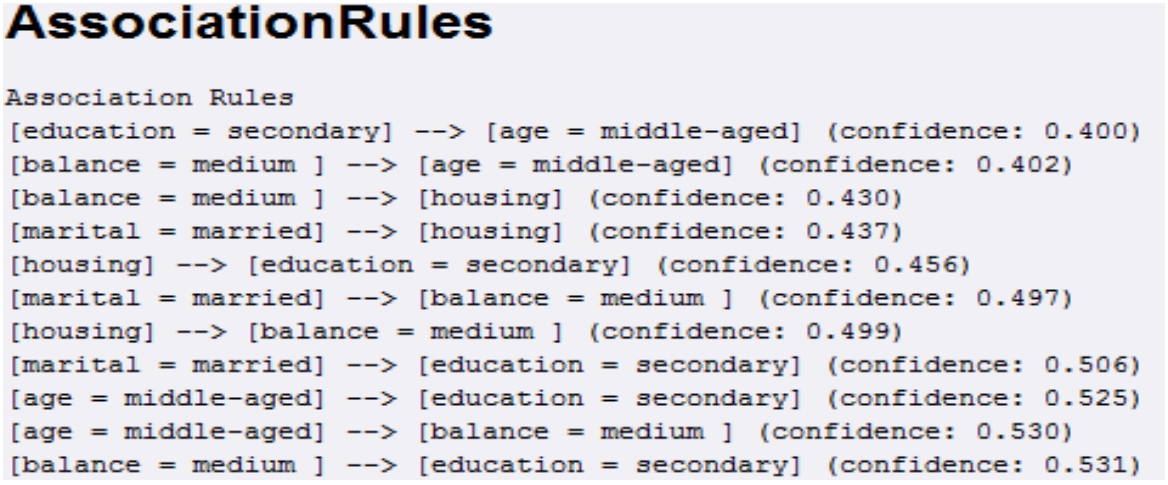


Figure-6 Association method output-Rapid Miner

Results of the Academic Research paper:

On paper [6]: According to data from a Portuguese bank, numbers of success out of 79345 contacts, which were made in 17 direct marketing campaigns between May 2018 and November 2010 using telephone, were 6499, meaning the success rate was 8.2%. For background information, the most used method of contact was through telephone, and as mentioned, a long-term, good interest rate deposit application was presented during each phone call. Although many assessment tools can be used, the most favored one for marketing campaigns is the cumulative Lift curve (Coppock 2002), that represents a percentage graph, in which each deciles represent a level or tier of, for example, responsiveness. In other words, a whole population is divided into deciles, and each member is placed into different deciles depending on its responsiveness; therefore, a group of the highest responder will be included in the first decile, and the lowest responder in the last decile. Such usage of Lift not only assess the accuracy and efficiency of the classifiers and models but also provides marketing managers with an efficient number of contacts that he or she can make.

Conclusion

In conclusion, the application of DM techniques has yielded better predictive performance and increased efficiency, as evident in three iterations of the Naive Bayes -DM methodology to the aforementioned data of a bank. It is manifested that the model that is materialized by a Rapid Miner tool and by SQL 2018 add-ins tool has a high predictive performance and practical knowledge that can assist managers in improving efficiency of campaign by directly addressing the issues: for example, changes in time period in which phone calls are made or length of phone calls made by interlocutors.

References

[1] Stefan M. Kostić, Miloš Đuričić, Mirjana I. Simić and Miroljub V. Kostić, "Data Mining and Modeling use Case in Banking Industry," in *Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 2018.

[2] L. I. Irina Ionita, "A Decision Support Based on Data Mining in e-Banking," in *Roedunet International Conference* , Romania, 2011.

[3] S. o. E. &. M. B. U. o. P. &. T. B. C. Shuxia Ren, "Customer segmentation of bank based on data warehouse and data mining," in *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*, 2010.

[4] H. Zhang, "The Optimality of Naive Bayes," in *International Journal of Pattern Recognition and Artificial Intelligence*, 2005.

[5] Ruguo Cao and Yongqin Tan, "Data Mining Program in library based on SQL Server 2005," in *2012 IEEE Symposium on Robotics and Applications (ISRA)*, Kuala Lumpur, Malaysia, 2012.

[6] P. C. Sérgio Moro and Raul M. S. Laureano, "USING DATA MINING FOR BANK DIRECT MARKETING:," in *Proceedings of the European Simulation and Modelling Conference*, Portugal, 2011.

[7] S. P. T. W. R. M. D. Twa, "Decision Tree Classification of Spatial Data Patterns," in *Decision Tree Classification of Spatial Data Patterns*.

[8] "Microsoft," [Online]. Available: http://msdn.microsoft.com/en-us/library/ms175595.

[9] K. Chitra Lekha and S. Prakasam, "Data mining techniques in detecting and predicting cyber crimes in banking sector," in *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, 2017.