Journal of Applied Data Sciences
Vol. 2, No. 3, September 2021, pp. 84-92

A Brief Overview of the Accuracy of Classification Algorithms for Data Prediction in Machine Learning Applications

Lichung Jen[1,*], Yu-Hsiang Lin[2]

[1] Department of International Business, National Taiwan University, Taiwan
[2] Department of International Business Administration, Chinese Culture University, Taiwan
lichung@management.ntu.edu.tw[1,*]; lyx21@ulive.pccu.edu.tw[2]
* corresponding author

(Received July 3, 2021 Revised August 12, 2021 Accepted August 28, 2021, Available online September 29, 2021)

Abstract

Many business applications rely on their history data to anticipate their company future. The marketing products process is one of the essential procedures for the firm. Customer needs supply a useful piece of information that helps to promote the suitable products at the proper moment. Moreover, services are recognized recently as products. The development of education and health services is reliant on historical data. For the more, lowering online social media networks problems and crimes need a big supply of information. Data analysts need to utilize an efficient categorization system to predict the future of such businesses. However, dealing with a vast quantity of data demands tremendous time to process. Data mining encompasses numerous valuable techniques that are used to anticipate statistical data in a number of business applications. The classification technique is one of the most extensively utilized with a range of algorithms. In this work, numerous categorization methods are revised in terms of accuracy in diverse domains of data mining applications. A complete analysis is done following delegated reading of 20 papers in the literature. This study intends to allow data analysts to identify the best suitable classification algorithm for numerous commercial applications including business in general, online social media networks, agriculture, health, and education. Results reveal FFBPN is the best accurate algorithm in the business arena. The Random Forest algorithm is the most accurate in categorizing online social networks (OSN) activity. Naïve Bayes method is the most accurate to classify agriculture datasets. OneR is the most accurate method to classify occurrences inside the health domain. The C4.5 Decision Tree method is the most accurate to classify students' records to forecast degree completion time.

Keywords: Data Prediction Techniques, Accuracy, Classification Algorithms, Data Mining Applications

1. Introduction

Decision-makers in the business industry are continually worrying about their corporate future. Since data collections comprise the primary resource of information, digitalizing company activities help to collect business operational data in enormous storages designated as a data warehouse. These past data can be used by data analysts to anticipate the future behavior of the business. However, dealing with a vast quantity of data demands tremendous time to process.

Data mining (DM) is a methodology that employs information technology and statistical approaches to search for prospective worthy information from a vast database that can be utilized to support administrative decisions. The rationale behind the relevance of DM is that data can be converted into usable information and knowledge automatically and intelligently. In addition, corporations employ data mining to know companies that work status and examine possible information values. Information gathered should be protected from the leaking of company secrets. Different data mining concepts were presented by Kaur [1] functions, material, and methods. Data mining is the use of sophisticated data analysis tools and procedures to find advanced ambiguities, patterns, and relationships that are valid in vast data sets. The best-known data mining approach is Association. In association, a pattern is discovered based on a relationship between goods in the same transaction. Clustering is a data mining tool that builds a useful

group of objects that have comparative qualities using the planned strategy. Decision Tree is one of the most common data mining approaches. One of the most difficult things to perform is when deciding to build a data mining framework is to know and decide which method to employ and when.

However, one of the most adopted data mining techniques in a variety of applications is the classification strategy. The classification procedure needs two types of data: training data and testing data. Training data are the data used by a data mining algorithm to learn the classification metrics to categorize the other data i.e. testing data. Many business applications rely on their history data to anticipate their company future. The literature covers many challenges that were solved by predicting through data mining approaches. In business, DM approaches are used to estimate the export abilities of firms [2]. In social media applications, missing link problems between online social networks (OSN) nodes are a frequent problem in which a link is expected to be between two nodes, but it becomes a missing link for several reasons [3]. In the agriculture industry, evaluating soil nutrients will show to be a big profit to the growers through automation and data mining [4].

Data mining approach is utilized to optimize the building energy performance through choosing the target multi-family housing complex (MFHC) for green remodeling [5]. In crime, preventing offense and force against the human female is one of the essential goals. Different data mining techniques were employed to examine the reasons of offense [6]. In the healthcare industry, numerous data mining algorithms have been applied to a range of diseases for detecting the infection in these diseases such as breast cancer detection, skin disorders, and blood diseases [7]. For the more, data analysts in the education industry employed data mining techniques to build learning strategies at schools and colleges [8]. Another goal is to detect several styles of learner behavior and forecast his performance [9]. One more goal is to forecast the student's pay after graduation based on the student's previous record and behavior during the study [10]. In general, services are considered products.

In this work, numerous categorization methods are revised in terms of accuracy in diverse domains of data mining applications. This study seeks to allow data analysts to choose the most suitable classification algorithm for numerous business applications including business, in general, lowering online social media networks difficulties, improving education, health and agriculture sector services. The present paper comprises of the following sections: Section 2 gives a methodology for numerous data mining approaches in the literature. Section 3 presents the results gained from the associated literature and further discussion. Finally, section 4 gives our conclusions and recommendations for further work.

2. Literature Review

The classification technique is one of the most implemented data mining techniques in a number of applications. The classification procedure needs two types of data: training data and testing data. Training data are the data used by a data mining algorithm to learn the classification metrics to categorize the other data i.e. testing data. Two data sets of text articles are employed and classified into training data and testing data. Three classic classification methods are compared in terms of accuracy and execution time by Besimi et al. [11]. K-nearest neighbor classifier (K-NN), Naïve Bayes classifier (NB), and Centroid classifier are explored. K-NN classifier is the slowest classifier since it uses the full training data as a reference to categorize testing data. On the other hand, the Centroid classifier employs the average vector for each class as a model to classify fresh data. Hence, the Centroid classifier is substantially faster than the K-NN classifier. In terms of accuracy, the Centroid classifier has the best accuracy rate among the others.

Several data mining techniques were utilized to forecast the export abilities of a sample of 272 enterprises by Silva et al. [2]. Synthetic Minority Oversampling Technique (SMOTE) is used to oversample unbalanced data. The K-means approach is used to group the sample into three separate groups. The generalized Regression Neural Network (GRNN) technique is used to reduce the error between the real input data points in the network and the regression predicting vector in the model. Feed Forward Back Propagation Neural Network (FFBPN) is a technique used in

machine learning to learn the pattern of certain input/output behavior for a set of data in a structure known as Artificial Neural Networks (ANN) (ANN). Support Vector Machine (SVM) is a classification technique used to classify a set of data according to similarities between them. A Decision Tree (DT) is a classification approach in which classes are expressed in a sequence of yes/no questions in a tree perspective. Naive Bayes is a classification approach used to classify one data set in numerous data sets according to the Bayes theorem probability idea. As a result, after using those strategies GRNN and FFBPN were the most accurate techniques utilized to predict the export abilities of enterprises.

Social media applications are developed based on Online Social Network (OSN) concept. Missing link problems between OSN nodes are a regular problem in which a link is meant to be between two nodes, but it becomes missing link due to some reasons. Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Decision Tree (DT), Neural Network, Naive Bayes (NB), Logistic Regression, and Random Forest are prediction approaches utilized to identify the missing link of two Facebook data sets by Sirisup and Songmuang [3]. One dataset (DS1) with high density and the other dataset (DS2) with low density. High density reflects that there is a high number of links between nodes. For high-density data collection, Random Forest delivers the greatest performance among the others in terms of accuracy, precision, F-measure, and area under the receiver operating characteristic curve (AUC) (AUC). On the other hand, the low-density data set can be predicted perfectly using either Random Forest or Decision Tree. In the conclusion, it can be said that Random Forest is the best prediction technique utilized to predict data in the OSN idea.

Analyzing soil nutrients will be proved to be a major profit to the growers. An agricultural survey has been leveraging on technical improvements such as automation, and data mining. Chiranjeevi and Ranjana [4] carried done a comparison examination of two algorithms i.e. Naive Bayes and J48. J48 is the improvement of the C4.5 classifier. A choice tree is a flowchart resembling a tree development, where each inner hub explains a test on a characteristic. Naive Bayes is a modest probabilistic classifier based on the Bayesian theorem with difficult naive individuality anticipation. Naive Bayes Algorithm can be tailored to foretell harvest growing in a soil specimen.

A decision support model was created for choosing the target multi-family housing complex (MFHC) for green remodeling using a data mining technique. Jeong et al. [5] locate the target of MFHC for green redesign that is important to establish a careful and intelligent evaluation approach of the building energy performance. The energy benchmark for MFHC in South Korea, but there was a drawback that the study was conducted on the MFHC used district heating system. To discover the green renovation aim of the MFHC, it is vital to respect different heating systems that are employed in MFHC e.g. individual heating systems, district heating systems, and central heating systems. However, there were two concerns about this study. First, the operational rating and energy benchmark system were proposed about the different variables of the heating system. Second, the model to discover the aim of MFHC for green redesign was produced regarding the diverse characteristics. The established decision support model can serve as a sensible benchmark to pinpoint the target of MFHC for green remodeling.

Preventing offense and force against the human female is one of the major tasks for police. Different data mining approaches were employed to evaluate the reasons of crime and the correlations between numerous offenses. These approaches play essential roles in offense analysis and forecasting. Kaur et al. [6] analyzes the data mining strategies utilized in offense predicting and analysis. It was concluded from this debate that most researchers used classification and clustering strategies for offense manner and disclosure. In the classification, the following approaches were used: Naïve Bayes, decision tree Bayesnet, J48, JRip, and OneR. For the more, Kumar et al. [12] proposed a data mining technique for cyber-attack difficulties. Many applications are incorporated in the cybersecurity idea. However, these apps need to be evaluated by data mining techniques to audit as a computer application. Deception of confidential information can arise through security crack access by an unauthorized person. Malicious software and viruses such as a trojan horse that is the reason for the infringement insecurity that leads to antisocial acts in the field of

cyber-crime. Data mining strategies that can be limited either secret information or data to legitimate users and unwanted access could be stopped.

However, Thongsatapornwatana [13] presents an overview of strategies used to investigate crime modes in past studies. The poll focuses on several forms of crimes e.g. violent crime, narcotics, border control, and cyber criminality. Survey results suggest that majority of the strategies used contain research gaps. These methodologies failed to accurately detect crime prediction, which increases the obstacles of overcoming this shortcoming. Hence, these strategies need criminal models, analysis, and prepare data to discover relevant algorithms. Data mining in the healthcare field is equally as crucial as studying diverse sectors. The aim of comprehending removal in health care records is a rigorous task and difficult. Mia et al. [7] study the different academic publications based on health care data to find the existing data mining methodologies and techniques provided. Many data mining technologies have been applied to a range of diseases for detecting the infection in these diseases such as breast cancer diagnosis, skin disorders, and blood diseases. Data mining execution has exceptional efficacy in this domain due to express amplification in the size of remedial data.

Moreover, Kaur and Bawa [14] give to the medical healthcare field a complete view of common data mining strategies to the researchers so that they can work more exploratory. Knowledge discovery in databases (KDD) examines vast volumes of data and turns it into relevant information. There is a boon to data mining techniques because it helps in the early diagnosis of medical conditions with high accuracy in which saves more time and money in any endeavor related to computers, robots, and parallel processing. Among all the medical ailments, cardiovascular is the most critical disease. Data mining is demonstrated efficacious as accuracy is a vital priority. Data mining techniques are demonstrated to be successfully employed in the treatment of numerous other major diseases which pose a threat to lives.

As another attempt, a comparison analysis is undertaken by Parsania et al. [15] to determine the best data mining classification approaches based on healthcare data in terms of accuracy, sensitivity, precision, false-positive rate, and f-measure. Naïve Bayes, Bayesian Network, J RIPPER (JRip), OneRule (OneR), and PART methods are selected to be used on a dataset from a health database. Results demonstrate that the PART approach is the best in terms of precision, false-positive rate, and f-measure metrics. In terms of accuracy, the OneR technique is the best whereas Bayesian Network is the greatest technique in terms of sensitivity.

Data mining techniques are used frequently in numerous fields. Data analysts in the education industry employed data mining techniques to build learning strategies at schools and universities since it serves a significant chunk of society. A corporative learning model to group learners into active learning groups through the web was presented by Amornsinlaphachai [8]. Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Naive Bayes (NB), Bayesian Belief Network (BN), RIPPER (called JRIP), ID3, and C4.5 (called J48) classification data mining algorithms are used to predict the performance of 474 students who study computer programming subject at Nakhon Ratchasima Rajabhat University in Thailand. A comparison of those algorithms is done to determine the most efficient algorithm among them.

As a result, C4.5 was the most efficient algorithm in forecasting students' academic performance levels in terms of several parameters such as correctness of the hypothesized data, data precision, recall, f-measure, mean absolute error, and processing time. Although C4.5 does not have the lowest processing time, it gets the highest percentage of correctness i.e. 74.89 percent as it is a simple and dependable algorithm. ID3 algorithm achieves the lowest proportion of correctness since its inception. Selecting learners to form active learning groups by the introduced model utilizing the C4.5 algorithm shows a better learning level vs traditional selecting by instructors.

To obtain a successful option that increases learner rendering and assists him to progress in schooling. Jalota and Agrawal [16] employed five classification approaches on the education dataset obtained through the Learning Management System (LMS) (LMS). Techniques that have been employed are the J48 algorithm, Support Vector

Machine algorithm, Naïve Bayes algorithm, Random Forest algorithm, and Multilayer Perceptron algorithm. All these technologies are beneath the Waikato Environment for Knowledge Analysis (WEKA) (WEKA). After comparisons, the results showed that Multilayer Perceptron outperformed other techniques since it received the top results in performance accuracy and performance metrics.

Roy and Garg [9] give a literature analysis of data mining approaches used in Educational Data Mining (EDM) (EDM). Data mining techniques are utilized in the EDM area to recognize distinct forms of learner behavior and forecast his performance. It was discovered that most of the prior studies collected data on predicting student performance through a set of questionnaires. The Cross-Industry Standard Process for Data Mining (CRISP-DM) model was employed. WEKA and (R tool) are data mining tools based on open-source language employed for statistical and data analysis.

As an application of data mining techniques in the education industry, Khongchai and Songmuang [10] devised an incentive for students by estimating the learner's future pay. Learners are often bored with academic studies. This can cause making their GPA poor or perhaps they leave college. It is due to the loss of motivation that motivates them to continue their education. To provide a suitable motivation for learners to make sure to continue their studies and develop their academic level. This can be achieved by suggesting a model that estimates the student's pay after graduation based on the student's previous record and behavior during the study.

In the meantime, the data mining techniques employed in this model are K-Nearest Neighbors (K-NN), Naive Bayes (NB), Decision trees J48, Multilayer Perceptron (MLP), and Support Vector Machines (SVM) (SVM). To establish the optimal technique for estimating future pay, a test was done by entering data of students graduating from the same university between the years 2006 to 2015. A WEKA (Waikato Environment for Knowledge Analysis) tool was used to compare the outputs of data mining approaches. The results revealed that after comparisons work outperformed (KNN) technique in predicting 84.69 percent for Recall, Precision, and F-measure. The other strategies were as follows: (J48) get a percentage of 73.96 percent, (SVM) (43.71 percent), Naive Bayes (NB) (43.63 percent), and Multilayer perceptron (MLP) (38.8 percent) (38.8 percent). A questionnaire was then issued to 50 current students at the university to assess if the model works to achieve its aims. The results of the questionnaire suggest that the proposed model boosted the motivation of the students and encouraged them to focus on continuing the study.

Sulieman and Jayakumari [17] proposed the importance of employing technology data mining 11th grade in Oman, which comprises a lot of units that supply the school in Oman administration inclusive student data. The goal is to lower the dropout rate of pupils and increase school performance. Using data mining techniques helps students to find the suitable maths for 11th grade in Oman. It is an opportunity to produce and deliver relevant analysis using such a system that extracts student information from the end-of-term grades to improve student performance. Knowledge derived from data mining enables decision-makers in the field of education make the appropriate option that will help in the development of educational processing. The math course incorporates data mining techniques. The findings of the numerous algorithms acquired from the various data using in a study that confirm the fact the prediction of student choice and performance may be produced using data mining techniques.

Academic databases used to be analyzed through a data mining approach to earn new valuable knowledge. Wati et al. [18] prophesy the degree-accomplishment time of bachelor's degree students by applying data mining technologies such as C4.5, and naive Bayes classification algorithms. They concentrate on the success of ranking data mining techniques specifically the C4.5 algorithm with its decision tree-based and naive Bayes classifier algorithm based on a gain ratio to discover the nodes. it demonstrates in the result of the anticipate degree achievement time of bachelor's degree the C4.5 algorithm is preferable in rendering gauge with (78 percent) precision, (85 percent) measured mean class precision, and (65 percent) measured mean class recall.

Anoop Kumar and Rahman [19] employed data mining techniques in inculcating a setting is called educational data mining (EDM) (EDM). The opportunities for data mining in education and the data to be reaped are illimitable.

Erudition discovered by data mining approaches may be utilized not only to utility the teachers to manage their courses and understand their students learning processes. As a result, all of these assists secure the advancement of kids in their academics and enforce few therapies if the progress is infeasible to the programming and institutional anticipation. The primary advantage is that kind of analysis avails to build a solution for sluggish learners. Useful for accomplishing educational data mining approaches which are using presently to advances in teaching and predict the performance of students to predict academic achievement in the learning process.

To conclude, strategies are utilized in data mining to change raw data to helpful reference in the education environment. Data mining in educational environments has extensive implementation. Educational environments result in a vast amount of student data, that is can be used for numerous objectives including forecasting the needs of students. Rambola et al. [20] compare the approaches and algorithms for data mining that are utilized in a different implementation, thereby rating their efficiency. Categorized aims of educational data mining can be realized in three types: prediction, clustering, and relationship mining. Some of the most frequent concepts, which are extensively employed in educational data mining, are described such as association rule mining, classification, clustering, and outlier detection rule. Association rule mining is applied for unsuccessful type extraction and to recommend the best path for the student.

3. Methodology

In this section, we outline the comparison results that were gathered from the literature in different business applications. Table 1 displays the comparison of classification algorithms that are used to predict data in business, online social media networks, agricultural, health, and education applications sectors.

Application Domain	Subdomain	Algorithms Used	Accuracy		Best Algorithm
General		K-NN NB Centroid Classifier	92.8 91.5 95.3		Centroid Classifier
Business		GRNN FFBPN SVM DT NB	83.3 85.2 77.8 70.8 72.2		Feed Forward Back Propagation Neural Network
Online Social Media Networks		SVM K-NN DT Neural Network NB Logistic Regression Random Forest	DS1 96 92 95 96 86 96 97	DS2 97 91 96 97 87 97 97	Random Forest
Agriculture		J48 DT NB	68 98		One Rule (OneR)

Education	Students Performance	Random Forest NB MLP SVM J48 DT	67.4 64.4 76.1 75.4 73.6	Multilayer Perceptron (MLP)
	Students Motivation	K-NN NB J48 DT MLP SVM	84.7 43.6 73.9 38.1 43.7	k-Nearest Neighbors (k-NN)
	Students Degree Completion	C4.5 DT	78	C4.5 Decision Tree (C4.5 DT)

As described in [11], k-nearest neighbors (k-NN) classifier, Naïve Bayes (NB) classifier, and Centroid classifier as classification techniques are compared. Politics, technology, and sports news stories are used with a total of 237 news articles. Experiments demonstrate that the Centroid classifier is the most accurate algorithm in classifying text documents since it classifies 226 news items accurately. Centroid classifier creates the average vector for each class and utilizes them as a reference to categorize each new test instance. However, k-NN needs to compare the test instance distance with all training instances distances for each time.

In [2], 272 companies are picked as a study sample to be categorized. Five classification techniques are used to categorize organizations into three classes: Generalized Regression Neural Network (GRNN), Feed Forward Back Propagation Neural Network (FFBPN), Support Vector Machine (SVM), Decision Tree (DT), and Naïve Bayes (NB) (NB). Results demonstrate that FFBPN is the most accurate algorithm to categorize instances in the business domain with an accuracy of 85.2 percent. Two Online Social Networks (OSN) datasets are used to compare the performance of seven classification algorithms. The first dataset (DS1) with High density (0.05) and the other dataset (DS2) with low-density (0.03). (0.03). The two datasets were collected using the Facebook API tool. Each dataset comprises public information about the people such as interests, friends, and demographics data. Classification techniques include; Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Decision Tree (DT), Neural Networks, Naïve Bayes (NB), Logistic Regression, and Random Forest. As results demonstrate in [3], the Random Forest method is the most accurate in categorizing OSN activities even with a high-density OSN dataset.

A collection of 1676 soil samples comprises 12 properties that need to be categorised. J48 Decision Tree (J48 DT) and Naïve Bayes (NB) classification methods are utilized. Results in [4] tells that the NB method is more accurate than J48 DT to categorize agriculture datasets since it classifies 98 percent of occurrences correctly. An experiment is undertaken in the health domain to classify 3163 patients' data as indicated in [15]. Naïve Bayes (NB), Bayesian Network (BayesNet), J Ripper (JRip), One Rule (OneR), and PART classification methods are utilized. Results demonstrate that OneR is the best accurate algorithm to categorize cases in the health domain with an accuracy of 99.2 percent.

Random Forest, Naïve Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and J48 Decision Tree (J48 DT) classification methods are utilized. 163 examples are utilized as an experimental dataset of students' performance. Results in [16] tell that the MLP algorithm is the best accurate algorithm to identify students' performance datasets since it identifies 76.1 percent of instances correctly. 13,541 students' profiles are utilized as a dataset to assess five categorization systems. k-Nearest Neighbors (k-NN), Naïve Bayes (NB), J48 Decision Tree

(J48 DT), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) were compared in terms of accuracy. As findings reveal in [10], the k-NN algorithm is the most accurate algorithm with an 84.7 percent accuracy level. 297 students' records were used as a dataset in [18]. Two classification techniques are applied: C4.5 Decision Tree (C4.5 DT), and Naïve Bayes (NB) (NB). Results suggest that the C4.5 DT algorithm is more accurate than NB to classify Students' data since it classifies 78 percent of occurrences correctly.

4. Conclusion

Data mining encompasses numerous valuable techniques that are used to anticipate statistical data in a number of business applications. The classification technique is one of the most extensively utilized with a range of algorithms. In this paper, various classification algorithms were revised in terms of accuracy in different areas of data mining applications including business in general, online social media networks, agriculture, health, and education to help data analysts to choose the most suitable classification algorithm for each business application. Experiments in the reviewed literature reveal that the Centroid classifier is the most accurate algorithm in classifying text texts. FFBPN is the most accurate algorithm to classify instances in the business domain. The Random Forest method is the most accurate in categorizing OSN activities. Naïve Bayes method is more accurate than J48 DT to classify agriculture datasets. OneR is the most accurate method to classify occurrences in the health domain. Multilayer Perceptron technique is the most accurate approach to categorize students' performance datasets. K-Nearest Neighbors algorithm is the most accurate algorithm in classifying students' profiles to boost their motivation. C4.5 Decision Tree method is more accurate than Naïve Bayes to classify students' records.

As future work, attention to review additional related papers in listed domains as well as uncover new areas will considerably add to the effort. Hence, the document will be utilized as a reference by business data analysts.

References

[1] Harkiran, K. (2017) A Study On Data Mining Techniques And Their Areas Of Application. International Journal of Recent Trends in Engineering and Research, 3, 93-95. https://doi.org/10.23883/IJRTER.2017.3393.EO7O3

[2] Silva, J., Borré, J.R., Castillo, A.P.P., Castro, L. and Varela, N. (2019) Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Export Potential of a Company. Procedia Computer Science, 151, 1194-1200. https://doi.org/10.1016/j.procs.2019.04.171

[3] Sirisup, C. and Songmuang, P. (2018) Exploring Efficiency of Data Mining Techniques for Missing Link in Online Social Network. 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, 15-17 November 2018. https://doi.org/10.1109/iSAI-NLP.2018.8692951

[4] Chiranjeevi, M.N. and Nadagoudar, R.B. (2018) Analysis of Soil Nutrients Using Data Mining Techniques. International Journal of Recent Trends in Engineering and Research, 4, 103-107. https://doi.org/10.23883/IJRTER.2018.4363.PDT1C

[5] Jeong, K., Hong, T., Chae, M. and Kim, J. (2019) Development of a Decision Support Model for Determining the Target Multi-Family Housing Complex for Green Remodeling Using Data Mining Techniques. Energy and Buildings, 202, Article ID: 109401. https://doi.org/10.1016/j.enbuild.2019.109401

[6] Kaur, B., Ahuja, L. and Kumar, V. (2019) Crime against Women: Analysis and Prediction Using Data Mining Techniques. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 14-16 February 2019, Faridabad. https://doi.org/10.1109/COMITCon.2019.8862195

[7] Mia, M.R., Hossain, S.A., Chhoton, A.C. and Chakraborty, N.R. (2018) A Comprehensive Study of Data Mining Techniques in Health-Care, Medical, and Bioinformatics. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 8-9 February 2018. https://doi.org/10.1109/IC4ME2.2018.8465626

[8] Amornsinlaphachai, P. (2016) Efficiency of Data Mining Models to Predict Academic Performance and a Cooperative Learning Model. 8th International Conference on Knowledge and Smart Technology (KST), Chiang Mai, 3-6 February 2016. https://doi.org/10.1109/KST.2016.7440483

[9] Roy, S. and Garg, A. (2017) Analyzing Performance of Students by Using Data Mining Techniques: A Literature Survey. 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 26-28 October 2017. https://doi.org/10.1109/UPCON.2017.8251035

[10] Khongchai, P. and Songmuang, P. (2017) Implement of Salary Prediction System to Improve Student Motivation Using Data Mining Technique. 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), Yogyakarta, 10-12 November 2016. https://doi.org/10.1109/KICSS.2016.7951419

[11] Besimi, N., Cico, B. and Besimi, A. (2017) Overview of Data Mining Classification Techniques: Traditional vs. Parallel/Distributed Programming Models. Proceedings of the 6th Mediterranean Conference on Embedded Computing, Bar, 11-15 June 2017, 1-4. https://doi.org/10.1109/MECO.2017.7977126

[12] Kumar, S.R., Jassi, J.S., Yadav, S.A. and Sharma, R. (2016) Data-Mining a Mechanism against Cyber Threats: A Review. International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Greater Noida, 3-5 February 2016. https://doi.org/10.1109/ICICCS.2016.7542343

[13] Thongsatapornwatana, U. (2016) A Survey of Data Mining Techniques for Analyzing Crime Patterns. Second Asian Conference on Defence Technology (ACDT), Chiang Mai, 21-23 January 2016. https://doi.org/10.1109/ACDT.2016.7437655

[14] Kaur, S. and Bawa, R.K. (2017) Data Mining for diagnosis in Healthcare Sector-a review, International Journal of Advances in Scientific Research and Engineering.

[15] T. wahyuningsih, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," J. Appl. Data Sci., vol. 2, no. 2, pp. 45–54, 2021, doi: 10.47738/jads.v2i2.31.

[16] T. T. Kim Phuong, "Proposing a Theoretical Model to Determine Factors Affecting on Job Satisfaction, Job Performance and Employees Loyalty For Technology Information (IT) Workers," Int. J. Appl. Inf. Manag., vol. 1, no. 4, pp. 201–209, 2021, doi: 10.47738/ijaim.v1i4.21.

[17] W.-J. Su, "The Effects of Safety Management Systems, Attitude and Commitment on Safety Behaviors and Performance," Int. J. Appl. Inf. Manag., vol. 1, no. 4, pp. 187–199, 2021, doi: 10.47738/ijaim.v1i4.20.

[18] Vaishali, S., Parsania, N., Jani, N. and Bhalodiya, N.H. (2014) Applying Naïve Bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis. International Journal of Darshan Institute on Engineering Research & Emerging Technologies, 3, 60-64.

[19] Jalota, C. and Agrawal, R. (2019) Analysis of Educational Data Mining using Classification. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, 14-16 February 2019. https://doi.org/10.1109/COMITCon.2019.8862214

[20] Al-Nadabi, S.S. and Jayakumari, C. (2019) Predict the Selection of Mathematics Subject for 11th Grade Students Using Data Mining Technique. 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, 15-16 January 2019. https://doi.org/10.1109/ICBDSC.2019.8645594

[21] Wati, M., Haeruddin and Indrawan, W. (2017) Predicting Degree-Completion Time with Data Mining. 3rd International Conference on Science in Information Technology (ICSITech), Bandung, 25-26 October 2017. https://doi.org/10.1109/ICSITech.2017.8257209

[22] Anoopkumar, M. andZubair Rahman, A.M.J.Md. (2016) A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration, International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, 16-18 March 2016.

[23] Rambola, R.K., Inamke, M. and Harne, S. (2018) Literature Review: Techniques and Algorithms Used for Various Applications of Educational Data Mining (EDM). 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, 14-15 December 2018. https://doi.org/10.1109/CCAA.2018.8777556