A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications

Ibrahim Ba'abbad, Thamer Althubiti, Abdulmohsen Alharbi, Khalid Alfarsi, Saim Rasheed

Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, KSA
Email: eabad@stu.kau.edu.sa, tmalthubiti@stu.kau.edu.sa, asalharbi0001@stu.kau.edu.sa, kayidhalfarsi@stu.kau.edu.sa, srahmed@kau.edu.sa

Abstract

Many business applications rely on their historical data to predict their business future. The marketing products process is one of the core processes for the business. Customer needs give a useful piece of information that helps to market the appropriate products at the appropriate time. Moreover, services are considered recently as products. The development of education and health services is depending on historical data. For the more, reducing online social media networks problems and crimes need a significant source of information. Data analysts need to use an efficient classification algorithm to predict the future of such businesses. However, dealing with a huge quantity of data requires great time to process. Data mining involves many useful techniques that are used to predict statistical data in a variety of business applications. The classification technique is one of the most widely used with a variety of algorithms. In this paper, various classification algorithms are revised in terms of accuracy in different areas of data mining applications. A comprehensive analysis is made after delegated reading of 20 papers in the literature. This paper aims to help data analysts to choose the most suitable classification algorithm for different business applications including business in general, online social media networks, agriculture, health, and education. Results show FFBPN is the most accurate algorithm in the business domain. The Random Forest algorithm is the most accurate in classifying online social networks (OSN) activities. Naïve Bayes algorithm is the most accurate to classify agriculture datasets. OneR is the most accurate algorithm to classify instances within the health domain. The C4.5 Decision Tree algorithm is the most accurate to classify students' records to predict degree completion time.

1. Introduction

Decision-makers in the business sector are always concerning about their business future. Since data collections form the core resource of information, digitalizing business activities help to collect business operational data in enormous storages named as a data warehouse. These historical data can be used by data analysts to predict the future behavior of the business. However, dealing with a huge quantity of data requires great time to process.

Data mining (DM) is a technique that uses information technology and statistical methods to search for potential worthy information from a large database that can be used to support administrative decisions. The reason behind the importance of DM is that data can be converted into useful information and knowledge automatically and intelligently. In addition, enterprises use data mining to know companies that work status and analyze potential information values. Information mined should be protected from the disclosure of company secrets.

Different data mining concepts were described by Kaur [1] functionalities, material, and mechanisms. Data mining involves the use of sophisticated data analysis tools and techniques to find advanced ambiguity, patterns, and relationships that are valid in large data sets. The best-known data mining technique is Association. In association, a pattern is discovered based on a relationship between items in the same transaction. Clustering is a data mining technology that creates a useful group of objects that have comparative features using the programmed strategy. Decision Tree is one of the most common data mining techniques. One of the most difficult things to do is when choosing to implement a data mining framework is to know and decide which method to use and when.

However, one of the most implemented data mining techniques in a variety of applications is the classification technique. The classification process 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 classify the other data *i.e.* testing data. Many business applications rely on their historical data to predict their business future.

The literature presents various problems that were solved by predicting through data mining techniques. In business, DM techniques are used to predict the export abilities of companies [2]. In social media applications, missing link problems between online social networks (OSN) nodes are a frequent problem in which a link is supposed to be between two nodes, but it becomes a missing link for some reasons [3]. In the agriculture sector, analyzing soil nutrients will prove to be a large profit to the growers through automation and data mining

[4].

Data mining technique is used to enhance the building energy performance through determining 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 important goals. Different data mining techniques were used to analyze the causes of offense [6]. In the healthcare sector, various data mining tools have been applied to a range of diseases for detecting the infection in these diseases such as breast cancer diagnosis, skin diseases, and blood diseases [7].

For the more, data analysts in the education field used data mining techniques to develop learning strategies at schools and universities [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 salary after graduation based on the student's previous record and behavior during the study [10]. In general, services are considered products.

In this paper, various classification algorithms are revised in terms of accuracy in different areas of data mining applications. This paper aims to help data analysts to choose the most suitable classification algorithm for different business applications including business, in general, reducing online social media networks problems, developing education, health and agriculture sector services.

The present paper consists of the following sections: Section 2 presents a methodology for several data mining techniques in the literature. Section 3 summarizes the results obtained from the related literature and further discussion. Finally, section 4 presents our conclusions and recommendations for future work.

2. Methods in Literature

The classification technique is one of the most implemented data mining techniques in a variety of applications. The classification process 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 classify the other data *i.e.* testing data. Two data sets of text articles are used and classified into training data and testing data. Three traditional classification algorithms 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 considered. K-NN classifier is the slowest classifier since it uses the whole training data as a reference to classify testing data. On the other hand, the Centroid classifier uses the average vector for each class as a model to classify new data. Hence, the Centroid classifier is much faster than the K-NN classifier. In terms of accuracy, the Centroid classifier has the highest accuracy rate among the others.

Several data mining techniques were used to predict the export abilities of a sample of 272 companies by Silva *et al.* [2]. Synthetic Minority Oversampling Technique (SMOTE) is used to oversample unbalanced data. The K-means method is used to group the sample into three different clusters. The generalized Regression Neural Network (GRNN) technique is used to minimize the error

between the actual 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 specific input/output behavior for a set of data in a structure known as Artificial Neural Networks (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 method in which classes are represented in a series of yes/no questions in a tree view. Naive Bayes is a classification technique used to classify one data set in several data sets according to the Bayes theorem probability concept. As a result, after applying those techniques GRNN and FFBPN were the most accurate techniques used to predict the export abilities of companies.

Social media applications are developed based on Online Social Network (OSN) concept. Missing link problems between OSN nodes are a frequent problem in which a link is supposed to be between two nodes, but it becomes missing link regard 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 techniques used to predict 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 huge number of links between nodes. For high-density data set, Random Forest gives the best performance among the others in terms of accuracy, precision, F-measure, and area under the receiver operating characteristic curve (AUC). On the other hand, the low-density data set can be predicted perfectly with either Random Forest or Decision Tree. In the end, it can be said that Random Forest is the best prediction technique used to predict data in the OSN concept.

Analyzing soil nutrients will be evidence to be a large profit to the growers. An agricultural survey has been capitalizing on technical advances such as automation, and data mining. Chiranjeevi and Ranjana [4] carried out a comparative analysis of two algorithms *i.e.* Naive Bayes and J48. J48 is the improvement of the C4.5 classifier. A choice tree is a flowchart like 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 tough naive individuality anticipation. Naive Bayes Algorithm can be bespoke to prophesy harvest growing in a soil specimen.

A decision support model was developed for determining the target multi-family housing complex (MFHC) for green remodeling using a data mining technique. Jeong *et al.* [5] locate the goal of MFHC for green remodel that is necessary to establish a careful and sensible evaluation method of the building energy performance. The energy benchmark for MFHC in South Korea, but there was a limitation that the study was conducted on the MFHC used district heating system. To locate the green remodel goal of the MFHC, it is necessary to regard different heating systems that are used in MFHC e.g. individual heating

systems, district heating systems, and central heating systems. However, there were two issues regarding this study. First, the operational rating and energy benchmark system were proposed regarding the different variables of the heating system. Second, the model to locate the goal of MFHC for green remodel was developed regarding the different characteristics. The developed decision support model can serve as a sensible standard to locate the goal of MFHC for green remodeling.

Preventing offense and force against the human female is one of the important goals for police. Different data mining techniques were used to analyze the causes of offense and the relationships between multiple offenses. These techniques play important roles in offense analysis and forecasting. Kaur *et al.* [6] reviews the data mining techniques used in offense forecasting and analysis. It was concluded from this discussion that most researchers used classification and clustering techniques for offense manner and disclosure. In the classification, the following techniques 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 issues. Many applications are included in the cybersecurity concept. However, these applications need to be analyzed by data mining techniques to audit as a computer application. Deception of secret information can happen through security crack access by an unauthorized user. Malicious software and viruses such as a trojan horse that is the reason for the infringement insecurity that leads to antisocial activities in the world of cyber-crime. Data mining techniques that can be restricting either secret information or data to legitimate users and unauthorized access could be blocked.

However, Thongsatapornwatana [13] provides a survey of techniques used to analyze crime modalities in previous research. The survey focuses on various types of crimes e.g. violent crime, drugs, border control, and cyber criminality. Survey results show that most of the techniques used contain research gaps. These techniques failed to accurately detect crime prediction, which increases the challenges of overcoming this failure. Hence, these techniques need crime models, analysis, and prepare data to find appropriate algorithms.

Data mining in the healthcare sector is just as important as exploring various areas. The mission of understanding removal in health care records is an exacting task and complex. Mia *et al.* [7] review the different academic literature based on health care data to find the existing data mining methods and techniques described. Many data mining tools have been applied to a range of diseases for detecting the infection in these diseases such as breast cancer diagnosis, skin diseases, and blood diseases. Data mining execution has high effectiveness in this domain due to express amplification in the size of remedial data.

Moreover, Kaur and Bawa [14] present to the medical healthcare sector a detailed view of popular data mining techniques to the researchers so that they can work more exploratory. Knowledge discovery in databases (KDD) analyzes large

volumes of data and turns it into meaningful information. There is a boon to data mining techniques because it helps in the early diagnosis of medical diseases with high accuracy in which saves more time and money in any effort related to computers, robots, and parallel processing. Among all the medical diseases, cardiovascular is the most critical disease. Data mining is proved efficacious as accuracy is a major concern. Data mining techniques are proved to be successfully used in the treatment of various other serious diseases which have a threat to lives.

As another attempt, a comparative analysis is conducted by Parsania *et al.* [15] to find the best data mining classification techniques 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 techniques are selected to be applied over a dataset from a health database. Results show that the PART technique is the best in terms of precision, false-positive rate, and f-measure metrics. In terms of accuracy, the OneR technique is the best while Bayesian Network is the best technique in terms of sensitivity.

Data mining techniques are used widely in several fields. Data analysts in the education field used data mining techniques to develop learning strategies at schools and universities since it serves a big chunk of society. A corporative learning model to group learners into active learning groups via the web was introduced 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 between those algorithms is made to select the most efficient algorithm among them.

As a result, C4.5 was the most efficient algorithm in predicting students' academic performance levels in terms of different measures such as correctness of the predicated 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 since it is a simple and reliable algorithm. ID3 algorithm gets the lowest percentage of correctness since its irrationality. Selecting learners to form active learning groups by the introduced model using the C4.5 algorithm shows a better learning level against traditional selecting by instructors.

To obtain a successful decision that improves learner rendering and helps him to proceed in education. Jalota and Agrawal [16] used five classification techniques on the education dataset collected through the Learning Management System (LMS). Techniques that have been used 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). After comparisons, the results showed that Multilayer Perceptron outperformed other techniques since

it got the highest results in performance accuracy and performance metrics.

Roy and Garg [9] present a literature survey of data mining techniques used in Educational Data Mining (EDM). Data mining techniques are used in the EDM domain to detect several styles of learner behavior and forecast his performance. It was concluded that most of the previous research collected data on predicting student performance by a set of questionnaires. The Cross-Industry Standard Process for Data Mining (CRISP-DM) model was used. WEKA and (R tool) are data mining tools based on open-source language applied for statistical and data analysis.

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

In the meantime, the data mining techniques used in this model are K-Nearest Neighbors (K-NN), Naive Bayes (NB), Decision trees J48, Multilayer Perceptron (MLP), and Support Vector Machines (SVM). To determine the preferable technique for predicting future salary, a test was conducted by entering data of students graduating from the same university during the years 2006 to 2015. A WEKA (Waikato Environment for Knowledge Analysis) tool was used to compare the outputs of data mining techniques. The results showed that after comparisons work outperformed (KNN) technique in predicting 84.69 percent for Recall, Precision, and F-measure. The other techniques 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). A questionnaire was then distributed to 50 current students at the university to see if the model works to achieve its objectives. The results of the questionnaire indicate that the proposed model increased the motivation of the students and helped them to focus on continuing the study.

Sulieman and Jayakumari [17] proposed the importance of using technology data mining 11th grade in Oman, which contains a lot of units that provide the school in Oman administration inclusive student data. The goal is to decrease the dropout rate of students and improve school performance. Using data mining techniques helps students to choose the appropriate mathematics for 11th grade in Oman. It is an opportunity to develop and give appropriate analysis through such a method that extracts student information from the end-of-term grades to improve student performance. Knowledge derives from data mining helps decision-makers in the field of education make the perfect decision that will help in the development of educational processing. The math subject uses

data mining techniques. The results of the various algorithms acquired from the various data using in a study that confirm the fact the prediction of student choice and performance can be obtained using data mining techniques.

Academic databases used to be analyzed through a data mining approach to earn new helpful knowledge. Wati *et al.* [18] prophesy the degree-accomplishment time of bachelor's degree students by using data mining algorithms such as C4.5, and naive Bayes classifier algorithms. They concentrate on the achievement of ranking data mining algorithms especially the C4.5 algorithm with its decision tree-based and naive Bayes classifier algorithm based on a gain ratio to find the nodes. it shows in the result of the foresee degree accomplishment 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] used data mining techniques in inculcating a setting is called educational data mining (EDM). The possibilities for data mining in education and the data to be reaped are illimitable. Erudition discovering by data mining techniques can be used not only to utility the teachers to manage their classes and understand their students learning processes. As a result, all of these helps ensure the advancement of students in their academics and enforce few treatments if the progress is infeasible to the programming and institutional anticipation. The basic advantage is that kind of analysis avails to establish a solution for slow learners. Useful for achieving educational data mining methods which are using presently to improvements in teaching and predict the performance of students to predict academic performance in the learning process.

To conclude, techniques are used in data mining to modify raw data to helpful reference in the education environment. Data mining in educational environments has widespread implementation. Educational environments result in a large amount of student data, that is can be used for different purposes like predicting the needs of students. Rambola *et al.* [20] compare the techniques and algorithms for data mining that are used in a different implementation, thus assessing their efficiency. Categorized objectives of educational data mining can be achieved in three types: prediction, clustering, and relationship mining. Some of the most common connotations, which are considerably used in educational data mining, are mentioned 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 course for the student.

3. Results and Discussion

In this section, we summarize the comparison results that were obtained from the literature in different business applications. Table 1 shows the comparison of classification algorithms that are used to predict data in business, online social media networks, agriculture, health, and education applications domains.

As mentioned in [11], k-nearest neighbors (k-NN) classifier, Naïve Bayes

Table 1. Comparison of classification algorithms in multiple applications.

Application Domain	Subdomain	Algorithms Used	Accuracy (percent)		Best Algorithm	Ref.
General (Text documents)		k-NN	92.8		Centroid Classifier	[11]
		NB	91.5			
		Centroid Classifier	95.3			
Business		GRNN	83.3		Feed Forward Back Propagation Neural Network (FFBPN)	[2]
		FFBPN	85.2			
		SVM	77.8			
		DT	70.8			
		NB	72.2			
Online Social Media Networks			DS1	DS2	Random Forest	[3]
		SVM	96	97		
		k-NN	92	91		
		DT	95	96		
		Neural Network	96	97		
		NB	86	87		
		Logistic Regression	96	97		
		Random Forest	97	97		
Agriculture		J48 DT	68		Naïve Bayes (NB)	[4]
		NB	98			
Health		NB	97.9		One Rule (OneR)	[15]
		BayesNet	98.5			
		JRip	97.5			
		OneR	99.2			
		PART	95.2			
Education	Students' Performance	Random Forest	67.4		Multilayer Perceptron (MLP)	[16]
		NB	64.4			
		MLP	76.1			
		SVM	75.4			
		J48 DT	73.6			
	Students' Motivation	k-NN	84.7		k-Nearest Neighbors (k-NN)	[10]
		NB	43.6			
		J48 DT	73.9			
		MLP	38.1			
		SVM	43.7			
	Students' Degree Completion Time	C4.5 DT	78		C4.5 Decision Tree (C4.5 DT)	[18]

(NB) classifier, and Centroid classifier as classification algorithms are compared. Politics, technology, and sports news articles are used with a total of 237 news articles. Experiments show that the Centroid classifier is the most accurate algorithm in classifying text documents since it classifies 226 news articles correctly. Centroid classifier calculates the average vector for each class and uses them as a reference to classify 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 taken as a study sample to be classified. Five classification algorithms are used to classify companies 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). Results show that FFBPN is the most accurate algorithm to classify 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). The two datasets were obtained using the Facebook API tool. Each dataset contains public information about the users such as interests, friends, and demographics data. Classification algorithms 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 show in [3], the Random Forest algorithm is the most accurate in classifying OSN activities even with a high-density OSN dataset.

A dataset of 1676 soil samples has 12 attributes that need to be classified. J48 Decision Tree (J48 DT) and Naïve Bayes (NB) classification algorithms are used. Results in [4] tells that the NB algorithm is more accurate than J48 DT to classify agriculture datasets since it classifies 98 percent of instances correctly.

An experiment is conducted in the health domain to classify 3163 patients' data as mentioned in [15]. Naïve Bayes (NB), Bayesian Network (BayesNet), J Ripper (JRip), One Rule (OneR), and PART classification algorithms are used. Results show that OneR is the most accurate algorithm to classify instances 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 algorithms are used. 163 instances are used as an experimental dataset of students' performance. Results in [16] tell that the MLP algorithm is the most accurate algorithm to classify students' performance datasets since it classifies 76.1 percent of instances correctly. 13,541 students' profiles are used as a dataset to examine five classification algorithms. 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 results show 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 classifi-

cation algorithms are applied: C4.5 Decision Tree (C4.5 DT), and Naïve Bayes (NB). Results tell that the C4.5 DT algorithm is more accurate than NB to classify Students' records since it classifies 78 percent of instances correctly.

4. Conclusions and Future Work

Data mining involves many useful techniques that are used to predict statistical data in a variety of business applications. The classification technique is one of the most widely used with a variety 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 show that the Centroid classifier is the most accurate algorithm in classifying text documents. FFBPN is the most accurate algorithm to classify instances in the business domain. The Random Forest algorithm is the most accurate in classifying OSN activities. Naïve Bayes algorithm is more accurate than J48 DT to classify agriculture datasets. OneR is the most accurate algorithm to classify instances in the health domain. Multilayer Perceptron algorithm is the most accurate algorithm to classify students' performance datasets. K-Nearest Neighbors algorithm is the most accurate algorithm in classifying students' profiles to increase their motivation. C4.5 Decision Tree algorithm is more accurate than Naïve Bayes to classify students' records.

As future work, consideration to review more related papers in mentioned domains as well as discover new domains will significantly add to the work. Hence, the paper will be used as a reference by business data analysts.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

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 Sup-

port 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. 8*th 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. 4*th 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. 11*th 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* 6*th 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] 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.

[16] 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

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

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

[19] 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.

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