

# Student Performance Prediction Using Case Based Reasoning Knowledge Base System (CBR-KBS) Based Data Mining

Prashant Dixit, Harish Nagar, and Sarvottam Dixit

**Abstract**—Higher education management problems in delivering 100% of graduates who can satisfy business demands. In industry it is often difficult for qualified graduates to identify the appropriate means to evaluate problem - solving abilities as well as shortcomings in the evaluation of problem solving skills. This is partially due to the lack of an adequate methodology. The purpose of this paper is to provide the appropriate CBR-KBS model for predicting and evaluating the characteristics of the student's dataset so as to comply with the parameters of selection required by the university industry. Machine learning algorithms have been used in these study areas under supervision, uncompleted and uncontrolled; K-Nearest neighbor, Naïve Bayes, Decision Tree, Neural Network, Logistic Regression and Vector Support Machines. The proposed model would allow university management to make easier, more professional, experienced and industry-specific plans for the manufacturing of graduates and graduates who passed the type I and II examinations held by the employment opportunities.

**Index Terms**—Case base reasoning, decision tree, knowledge base system, neural network, WEKA.

## I. INTRODUCTION

Graduates have raised their rate of unemployment from one year to the next. A total of 200,000 university graduates graduated annually. Six months after graduation 1 out of 4 graduates are unemployed and the rest are graduates. It is very difficult to know how qualified graduates appear for college, even though this field is not explored [1]. A research is also needed to develop a successful understanding of the industry's needs. In the light of the disturbing patterns in graduate unemployment we must start thinking, researching, planning, investigating, constructing and implementing some kind of tools to evaluate employability to help us correct shortcomings or Improve academic results after graduation. Many surveys have been conducted to identify the pattern in graduates' productivity and job status [2]. One of these is the analysis of the graduate tracer by the higher education department. This research examined the job situation of fresh graduates who live, study or still actively pursue jobs prior to the season. Researchers have recorded the findings of tackling the problem of employee turnover. The Ministry of Higher Education produced a total of 107 850 bachelor

respondent who took part in the 2015 online survey. The findings found 58.0% of graduates who were working at first, while 27.9% were still unemployment, 6.2% were looking for jobs, 5.6% and 2.2%, and additional experiments continue to improve their skills. The survey conducted by Kementerian Pengajian Tinggi in 2014 included 88.7% working in full time and 11.3% worked in part-time.

According to information from the Ministry of Higher Education, a total of 28,148 people have graduations and their proportion have not been at public university, compared with 52,219 graduates in their entirety from the group of institutions of higher education; (HEIs). This is a very worrying figure, given that the results suggest that the highest unemployment rate is for unemployed graduates from public universities. University graduate unemployment can also be classified by area of study. It was also shown that the highest unemployment was in the art and social sciences in 2014. The education sector, however, has also shown that the unemployment rate rose from 22 percent in 2014 to 22.1 percent in 2013.

According to the report, some 64% of employers said they don't care whether student from universities abroad, private or public. There were three main reasons for unemployment among new graduates, approximately 64% of new graduates were under control in English, 60% were poor in communications skills and 59% were poor in attitudes [3]. Around 88% of employers reported that they were keeping or growing their employment in 2016. These figures suggest that the electoral staff are better suited to and have a range of qualifications and benefits, in particular with regard to adaptability. Jake David said that applicants should acquire transferable qualifications, not just specialists, to allow them to play a different job roles, according to SEEK Asia Chief Officer. The qualities acquired are translatable, such as interpersonal skill, listening skills, management abilities and leadership skills that are transferred to the job world [4]. Examples of organizational skills include support for people, dispute resolution and internally motivated inspiration, while example for communications include public speaking, consulting, convincing or selling. Objectives, time management and multifaceted work are examples of interpersonal skills, while leadership skills are examples of decision-making, motivation and group working. Most companies see that highly qualified workers have earned amazing personalities the employer's favor. A new study by JobStreet.com shows that just 14 percent of employers prefer skills, while most of the 51 percent of employers consider personalities as the main selection criterion. Six main factors for new graduates are 59% of employers who say that they

Manuscript received May 27, 2021; revised July 20, 2021.

Prashant Dixit and Harish Nagar are with Sangam University, Bhilwara, Rajasthan, India (e-mail: prashantfpc@gmail.com, harishngr@gmail.com).

Sarvottam Dixit is with Mewar University, Rajasthan, India (e-mail: sdixit\_dr@rediffmail.com).

have low attitude or character among graduates, 64% poor English order, 60% limited vocabulary and difficulties with skills, decision making and problem-solving [5]. The analysis connects the data set's characteristics with all six capabilities needed to fulfill the employer's requirements. So, if statistical modeling and data mining techniques may allow different parties to schedule postgraduate jobs. Data mining is a critical procedure for finding a vast number of data patterns and information. Data mining was known for the use of Facebook, Microsoft, Google and more in technology companies around the world. By data processing the data mining process is about describing the past and forecasting the future. Database infrastructure, statistics, visualization techniques and machine intelligence are all rooted in data mining. There are several activities included: Overview definition, study of associations, classification and estimation, analyzes of clusters, outlier analysis, predictive analysis, analysis of trends and assessments, regression, etc [6]. The method of development of a model to forecast the outcome is predictive modeling. When the result is categorical, so grouping is called, but if the result is empirical, regression is called. Clustering is assigned in description to cluster observations so that they are identical in the same cluster and association rules may find interesting correlations within hypotheses. The most critical components of data mining are classification and forecasting. The algorithm for selection also depends on the type of information to be used (nominal, ordinal, ratio or interval). Any data mining algorithm is provided with a machine learning category and different data mining algorithms are used to set knowledge-based data [7].

## II. LITERATURE SURVEY

One of the main tasks of data mining that researchers have used extensively in the field of prediction science are the supervised learning technologies. For experiments on widely used model of graduate employment, classification and prediction such as Naive Bayes, the Neural Network, Decision Tree, Logistic Regressions, Random Forest, etc. WEKA also employs as many data processing techniques to build statistical models for the defect prediction. A prediction research performed with KDD (Database knowledge discovery) and CRISPDM (Cross Industry Norm Data Mining Procedure) as study methods. Improve academic results after graduation. Several surveys have been conducted to identify the pattern in graduates' productivity and job status. One of them is the analysis of the graduate tracer by the higher education department. This research examined the job situation of fresh graduates who live, study or still actively pursue jobs prior to the season. Researchers have documented the findings of tackling the problem of employee turnover. The Ministry of Higher Education produced a total of 107 850 bachelor participants who took part in the 2015 online survey.

The Graduate Employability model was developed by Shahiri and Husain [8]. This research was using the Tracer Study data carried out by the Malaysian Higher Education Ministry covering both polytechnic, public and private graduates. During the first six months of graduates jobs, unemployment, or conditions that cannot be decided in the

first six months after the graduating year, used Bayesian and Decision Tree algorithms. The findings of the analysis find that classification J48 of a decision-making arboreal variant is as accurate as Classifier.

Bunkar [9] used five methods for data mining; Naive Bayes, Logistic Regression, Multilayer Perceptron and J48 decision-making tree. Data from the Examining Units, Alumni Units and Research Units. Data gathered. The findings reveal that the logistic regression is the best classification to predict whether graduates serve or pursue their education in private or public sectors.

In order to forecast student results, Kaswan *et al.* [10] used five classification tasks such as ANN, Naive Bayes, KNN, SVM, and Decision Tree. Demographics, internal and external assessment are the attributes used for this analysis. The findings of the experiment shows that Neural Network's prediction error is smaller and more accurate than other tasks.

Kaswan *et al.* [11] suggested a student success prediction method used the WEKA Naive Bayes method for building prediction models. Six properties have been tested: income households, university, ethnicity, gender, home town and CGPA. There are three of them. The study finds that the income of the sex, the hometown and the family contributes to university success.

The Bayesian Methods used by Kaswan *et al.* [12] comprise Naive Bayesian simple, Naif Bayesian, Averaged One Dependence Estimators (ADES), ADSE, Bayesian networks, and Naif Bayesian update. The results revealed an average one-dependence approximation, with an algorithm for subsumption resolution (AODEsr) of 98.3%, followed by the 96.1% for AODE.

In order to analysis the work using WEKA Gao construct a data mining model. The parameters are Jobs Data. For evaluation by analyzing and comparing based on various parameters, Classical Decision Tree classifiers were used. Three results can be drawn which are now almost half of the students who do not choose the school, there is a connection between the origin of the position of students and each task for each gender is unique.

Kaswan *et al.* [13] has carried out research which has developed an employment model for graduates that could forecast whether they are unemployed, are working or are not in a state. The contrast of the Bayesian and the Decision treaties 12 months after completion, the study defined qualities which may influence graduates' jobs. The report used a survey of 11,853 samples from the Maejo University Planning Office in Thailand. To construct the classification, merged the process of the discovery of information, and the cross-industry provides the following for data mining. The study showed that Waode of Bayesian methods was the highest accuracy of 99.77%.

Arsad and Buniyamin [14] conducted research to forecast the positioning of the student in Work Integrated Learning (WIL). In this study students were classified using J48, Bayes Net, Naive Bayes, mobile cart and REPTREE algorithm. The characteristics of this r include sex, participation, supporters, subjects and semester grades the findings revealed that the Bayes Net and Naive Bayes algorithms provide strong results and effectively estimate how many students have passed or failed WIL. In general, Decision Tree, Naive Bayes, Logistic

Regression, Network of Bayes, and Averaged One-Dependence Estimator (AODESR) chosen by researchers as a statistical model due to their best experimental reliability.

### III. CASE BASE REASONING

Case-based learning is an AI methodology that resolves issues by comparing certain related problems (past experience) previously solved, held in a case basis, to a current unresolved problem to draw analogical conclusions to the problem resolution [15]. The AI approach providing an artificial intelligence approach is a problem-solution approach. Problems solving CBR generally requires the retrieval of applicable prior, related cases and the adaptation, if possible, of a solution(s) to the prior case(s), to resolve the new problem and to save the present case in the future as a new case. It was defined in different ways by various scholars, but the main aim of the technique was the same. "The case-based reasoning is both the way people use cases to solve problems and the manner in which machetes are used to solve problems," defined Riesbeck and Shank (1989) as "Case-based reasoning is both the way some people utilize cases to fix problems and the manner in which they may be used" CBR is both a problem solving approach and a thinking and perception philosophy that is focused around how people understand the memories of the past. It brings together elements of both KBSs and the world of master learning. In recent decades, the CBR study has flourished in a range of domain areas. However, it provides a systemic means to resolve problems regardless of the field of implementation. The key phases of the CBR scheme are in given fig:

- the input issue is characterized by proper function indices (attributes)
- the memory may contain related cases relative to the current situation.
- adaptation to current issue of a prior approach or a series of solutions
- assessment of the modified solution
- storage of the appropriate solution or the experience analyzed.

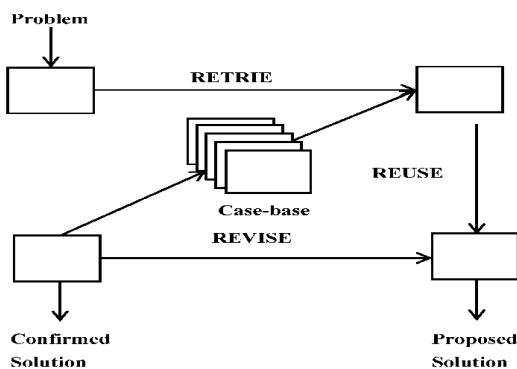


Fig. 1. Framework of case base reasoning.

### IV. KNOWLEDGE BASE SYSTEM

Every level of information – syntactic, semantics, logic, budget, comprehension of unformed input, ellipticity, cases constraints, vagueness – must be represented by the KR

scheme [16] The K-Box, the Knowledge-Base, the query applicator, logic, and the user interface are split into five sections to make this model more efficiently.

#### A. K Rectangle

The first portion of the K Rectangle contains the external input from the user interface. A text, a novel, a paper etc are the source of the information. Input can be categorized into: nature, quality, behavior, generality, specificity, inherence, falsity and each material can be earth, water, sun, air, atmosphere, period, space, mind and soul as is the case with Indian logic. The user's input can either be a new knowledge or the question is split into two groups. If a person asks for an examination whether a database of facts is like solving an inquir.

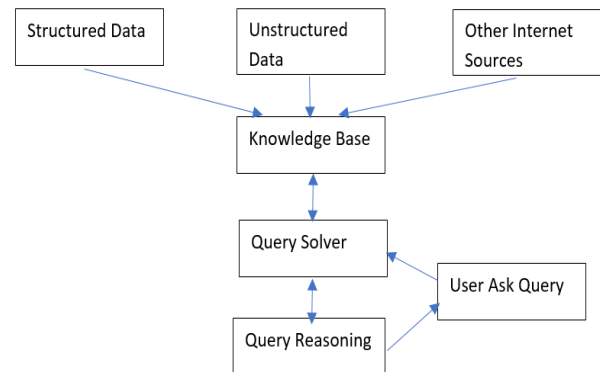


Fig. 2. Framework of knowledge base.

When the new material is entered, it goes through the acquisition and learning phase, to verify if the knowledge is already in the knowledge base. Otherwise, it verifies that the current procedure takes account of such information if a segmenting procedure has been carried out on the input to verify which groups it falls within and distinguishes the operation from the other. The K Package Feature Extraction component can verify whether a method can be performed or whether the received text is present for Ex. Mobile rings, so the step of this new knowledge takes place, means that a sound is arriving and the source of the ringing is being played. If the statement is like ram, so no action is taken. If we use somaticized net, framing and logic programming, but if we can do any operation by the object, then a structure is needed that can be complex and need to be articulate. In this way, the information becomes a simple phrase. The Information Structure component of the K box is used by the best knowledge representation technology to describe incoming knowledge. Semantic Net and Script methods are combined with KR. All the expertise to solve the dilemma is a body of knowledge. The basis of information may be general or unique domain. Want to do it? The applicator is used to obtain the information from the device and then transfers the information to the logic process. Whenever the latest query enters a method, it will discover whether the query is identical to the earlier query or generating from the responsibility strategy. At this point we can make the machine intelligent with the association of learning rules mining. Reasoning is used to obtain additional information from current experience. Forward and backward logic is the easiest method of thought.

V. RESULT AND DISCUSSION

A. Student Selection Criteria

The working flow of the classification data mining model. In the initial stage, the pre-exam data score is 80% followed by key exam function classification methods. Then classification methods are used to create a successful flow that can translate inputs correctly to desired outputs. Feedback on the collection of features and the learning phases to increase classification efficiency are provided during the model evaluation process. When a template is created, in the second step, the student test results are predicted to have a mark of 90 percent and are qualified for work .

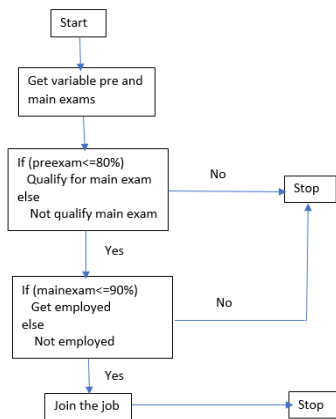


Fig. 3. Flowchart student selection criteria.

B. With the Help of CBR-KBS for Selection for Employment

One of the positive stories of Ai technologies is the expert or KBSs. The KBSs are built on an explicitly process of information representation needed to solve a problem - so called Second Generation systems which make it possible for a system to reasoning using the rule-based reasoning theory (RBR). RBR is also developing some specialist programs. Yet, regardless of whether the information is profound or superficial, an explicit domain model still has to be created and sometimes applied as rules or maybe more recently as an object model. There are undoubtedly several issues in model-based KBS, but there are many. The challenges are:

- the acquisition of knowledge is a complex process that is often called the bottlenecks for eliciting knowledge
- the implementation of KBS requires specific expertise which often takes long years
- model-based KBS often slows down and does not make it easy for them to reach and handle vast quantities of information.

Several attempts were made to resolve the problems by improving exploration and mining techniques and methods, improving and practical KBS shells and ecosystems, improving implementation methodologies, language and ontology modeling of knowledges, promoting collaboration between KBS and the database in expert and inductive reasoning databases, and systems-keeping techniques and means. In recent years, however, it is necessary to create a straightforward human reasoning model and a framework for solving nonlinear equations that leads to the advancement of case-basis reasoning. In certain ways, CBR differs radically

from other big AI methods because it does not depend on the specific physical equation describing a problem's domain information, but uses general relationships between its characteristics. This will resolve problems with previous KBSs such as:

- CBR does not have to have explicit domain patterns and hence elicitation becomes a task to collect case history
- Implementation reduces to recognizing important features describing case studies; it's simpler to do this than to create an integrative framework.

This CBR-KBS presents us with all specifics of the applicant residing in the CBR knowledge base to pick the right applicant for a good job opportunity. In future, as CBR keeps record in its database, as the case may be, in the below image.

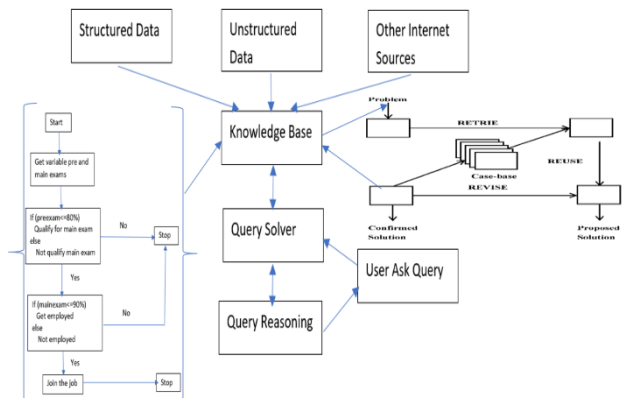


Fig. 4. Framework of CBR-KBS for student selection for employment.

C. Training Data Sets

New dataset is used for building proposed model. For training we have 2159 both (Pre-exam Type I) instances of students' data with their academic and industrial exam performance from pre-exam (type A), aptitude exam and main-exam (type B) courses related exam, number of question, evaluate correct and wrong question and scoring the result presented information in Table I.

TABLE I: DESCRIPTION OF DATA SETS

Features	Description
UID	Student Unique Identification (Roll No)
Name	Name of student
Gender	Male, Female
Course	Course ID
Subject	Subject name and subject code
Exam duration	Mid-term (MT), Final-term (FT)
Question	Number of Question
Correct	Correct answer
Wrong	Wrong answer
Total score	Gain percentage
Psychometric Factor	To evaluate student marks table II and III

D. Sample Data

CBR-KBS techniques have been used to build a classifier that can predict the performance of the students in an economics subject that is offered by Orta Dogu Teknik Universitesi (<https://data.mendeley.com>). Table II consists of training data set of ten students. Input data for the classifiers are provided from CSV file includes student's marks from pre exam, data regarding obtained marks in pre-exam, exam type, student ID, gender, pre-exam type details. Table II: The

sample input data. Every input data is in alphanumeric form. As in Table II, 16 students selected in green color cleared pre exam out of 29 students appeared in pre examination

TABLE II: PRE-EXAM RESULT

ID-Scrambled	Gender	Pre-Exam Type	Correct	Wrong	Score
2438768	F	A	23	2	92
2568626	M	A	16	9	64
2436278	F	A	19	6	76
2437482	F	A	21	4	84
2705146	F	A	20	5	80
2303670	F	A	23	2	92
2293686	M	A	18	7	72
2999998	M	A	12	13	48
2438950	F	A	24	1	96
2435074	F	A	11	14	44
2436324	F	A	20	5	80
2580968	F	A	10	15	40
2303700	M	A	18	7	72
1215030	F	A	21	4	84
2436392	F	A	23	2	92
2648400	F	A	22	3	88
2244166	M	A	16	9	64
2436440	F	A	20	5	80
2217092	F	A	16	9	64
2435156	F	A	21	4	84
2439180	F	A	24	1	96
2442942	M	A	20	5	80
2436492	F	A	21	4	84
2217122	F	A	20	5	80
1855878	M	A	12	13	48
2581166	F	A	14	11	56
2581218	F	A	15	10	60
2307728	M	A	21	4	84
2581214	F	A	19	6	76

Fig. 5 show the pre-exam result whose candidate score 80% or more they qualify mains but less they not qualify.

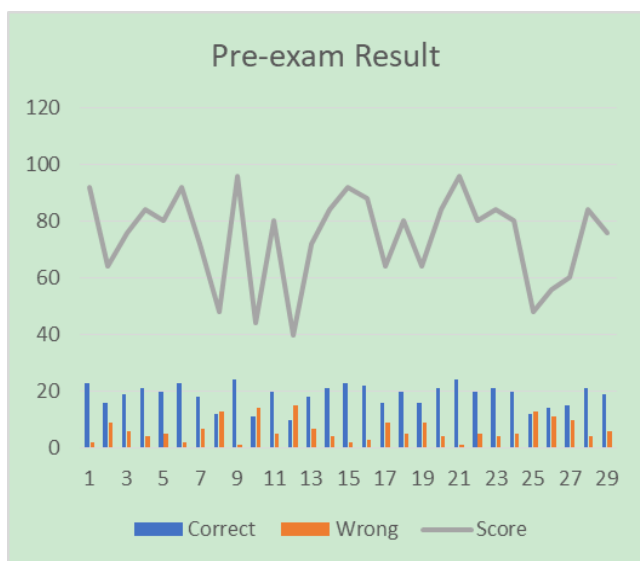


Fig. 5. Scoring results in pre exam.

In Table III, 7 students selected in blue color cleared main exam out of 16 students appeared in main examination and get employed in industry

TABLE III: MAIN-EXAM RESULT

D-Scrambled	Gender	Main-Exam Type	Correct	Wrong	Score
2438768	F	A	21	4	82
2437482	F	A	23	2	90
2705146	F	A	18	6	79
2303670	F	A	23	2	91
2438950	F	A	24	1	97
2436324	F	A	16	9	85
1215030	F	A	15	10	82
2436392	F	A	24	1	96
2648400	F	A	18	7	86
2436440	F	A	23	2	95
2435156	F	A	11	14	70
2439180	F	A	24	1	96
2442942	M	A	21	4	93
2436492	F	A	15	10	81
2217122	F	A	14	11	71
2307728	M	A	21	4	77

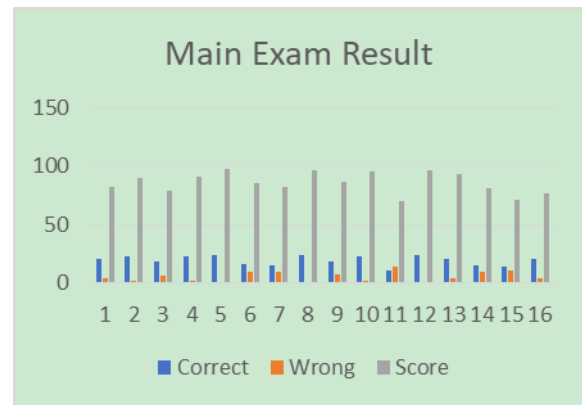


Fig. 6. Scoring results in main exam.

Fig. 6 shows the main-exam result whose candidate score 90% or more they qualify mains but less they not qualify and liable to get the employment

## VI. CONCLUSION

The success forecast is established in formal education and has been contributed by several scholars. There is, however, no study into the estimation of employee engagement. Within the scope of the institution's success and employment opportunities, analysis is expected to create robust performance and employment opportunities models and a method capable of predicting performances and employment opportunities. In this article, CBR-KBS techniques were explored and these techniques were very beneficial to pick the best student for jobs.

## CONFLICT OF INTEREST

I have collected data from this link (<https://data.mendeley.com>). It is available and freeware on google data set. No any type of conflict generate in future.



#### AUTHOR CONTRIBUTIONS

Prof. Sarvottam Dixit and Dr. Harish Nager deep knowledge in machine learning. They have imply dataset in ML techniques to get better results to learn Mr. Prashant Dixit from them to use their learning techniques than to write this paper.

#### ACKNOWLEDGMENT

I highly thanks to google data sets and my guide Prof. Sarvottam Dixit and Dr. Harish Nagar they help me in this research paper.

#### REFERENCES

- [1] J. C. Felizardo, K. R. Macdonell, S. G. Mendes, and E. Maldonado, "A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews," arXiv Prepr., arXiv:2012.10833, 2020.
- [2] M. Liljeström, H. Paulin, and C. Holmgren, "In search of online learning 3.0 in higher education—the role of theory-based process evaluation," in *Proc. ICICTE2019, International Conference on Information Communication Technologies in Education*, 2019, pp. 21–30.
- [3] T. K äser, N. R. Hallinen, and D. L. Schwartz, "Modeling exploration strategies to predict student performance within a learning environment and beyond," in *Proc. the Seventh International Learning Analytics & Knowledge Conference*, 2017, ACM.
- [4] C. Marquez-Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *Tecnologías del Aprendizaje, IEEE Revista Iberoamericana de*, 2013, vol. 8, no. 1, pp. 7–14.
- [5] S. Rovira, E. Puertas, and L. Igual, "Data-driven system to predict academic grades and dropout," *PLoS One*, 2017, vol. 12, no. 2, p. e0171207.
- [6] M. Awad, K. Salameh, and E. L. Leiss, "Evaluating learning management system usage at a small university," in *Proc. the 2019 3rd International Conference 260 on Information System and Data Mining*, 2019, pp. 98–102.
- [7] R. Mendes and J. P. Vilela, *Privacy-Preserving Data Mining: Methods, Metrics, and Applications*, IEEE, 2017.
- [8] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, 2015, vol. 72, pp. 414–422.
- [9] K. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification," in *Proc. 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, 2012, IEEE.
- [10] K. S. Kaswan, "A new nature inspired optimization technique," *International Journal of Advanced Research Engineering and Technology (IJARET)*, vol. 12, issue 2, 2021.

- [11] S. Sharma, N. Kumar, and K. S. Kaswan, "Big data reliability: A critical review," *Journal of Intelligent & Fuzzy Systems*, vol. 41, pp. 1–16, 2021.
- [12] V. Bhatia, V. Jaglan, S. Kumawat, and K. S. Kaswan, "A hidden markov model based prediction mechanism for cluster head selection in WSN," *International Journal of Advanced Science and Technology*, vol. 28, no. 15, 2019, pp. 585–600.
- [13] K. S. Kaswan, "Fault model for UML behavioral activity and sequence diagrams," *International Journal of Engineering and Advanced Technology (IJEAT)*, ISSN: 2249-8958, vol. 9, issue 1, October 2019.
- [14] P. M. Arsal, N. Buniyamin, and J.-I. A. Manan, "A neural network students' performance prediction model (NNSPPM)," in *Proc. 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2013, IEEE.
- [15] S. Srinivasan, J. Singh, and V. Kumar, "Multi-agent based decision support system using data mining and case based reasoning," *IJCSI International Journal of Computer Science Issues*, vol. 8, issue 4, no. 2, July 2011, ISSN (Online): 1694-0814.
- [16] J. S. Dhatwal, S. Dixit, and S. Srinivasan, "Implementation of case base reasoning system using multi-agent system technology for a buyer and seller negotiation system," *International Journal of Electronics and Communication Engineering*, 2019.

Copyright © 2022 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).



**Parshant Dixit** is a PhD student in the Department of Computer Science and Engineering, Sangam University. He earned his B.Tech from Uttar Pradesh Technical University, Lucknow and M.Tech degrees from Sunrise University, Alwar, Raj India. His research interests include artificial intelligence and agent technology.



**Harish Nagar** completed his PhD from the University of Kota. Presently, he is a professor at the Sangam University, Bhilwara, Rajasthan, India. His research interests include artificial intelligence, machine learning, software engineering etc.



**Sarvottam Dixit** earned his PhD from Agra University, India in 1990. Currently, he is serving as the professor and advisor to chancellor and pro vice chancellor at Mewar University, India. His general research interests include artificial intelligence, machine learning, multi agent etc.