Survey on Assessing Students’ Performance using Data Mining Techniques

Shaimaa Mohsen Hassan
Department of Information Systems and Technology Faculty of Graduate Studies for Statistical Research Cairo University Egypt

Nesrine Ali AbdelAzim
Department of Information Systems and Technology
Faculty of Graduate Studies for Statistical Research Cairo University Egypt

Nagy Ramadan
Department of Information Systems and Technology Faculty of Graduate Studies for Statistical Research Cairo University Egypt

ABSTRACT

Finding patterns and other relevant information from huge datasets is a process known as data mining, also referred to as Knowledge Discovery in Database (KDD). Large data sets are sorted through in data mining in order to find patterns and relationships that may be used in data analysis to assist solve business challenges. Enterprises can forecast future trends and make more educated business decisions due to data mining techniques and technologies. Data science uses cutting-edge analytics techniques to find important information in data sets, and data mining is an essential part of data analytics as a whole and one of the core areas of data science. At a more granular level, data mining is a step in KDD process, a data science methodology for gathering, processing and analyzing data [31]. Data Mining techniques can be used in educational field which is called Educational Data Mining; it is an emerging discipline concerned with developing methods. These methods are used for exploring unique and increasingly large-scale data that come from educational settings in order to better understand students, and the settings which they learn in. [33]. This study presents a literature survey upon the various data mining tasks used in the prediction of the students’ performance.

General Terms

Data Mining, Educational Data Mining.

Keywords

Classification, Association Rules, Students’ Performance.

1. INTRODUCTION

The growing volume of data in educational systems makes it difficult to predict students' success. Educational Data Mining (EDM) field will help us to predict the low performing students early enough to overcome their difficulties in learning and improve their learning outcomes, which in turns serves the institutional goals of providing high quality education system. Data mining is the process of analyzing data from different perspectives and summarizes it into useful information. Finding connections or patterns between fields in a relational database is what is referred to as data mining technically. Recently, both the amount of student digital data and the use of online technology in education have increased. As a result, it is now possible to use data mining techniques to process educational data and develop predictions and rules for the pupils. All kinds of information about the student’s socioeconomic environment, learning environment, or course grades can be used for prediction, which affect the success or failure of a student [3]. The most common methods are Supervised learning and Unsupervised learning. The function that is learned from the training portion of the dataset is the subject of supervised learning. A supervised learning method uses the training data that is already available to create an inferred function that can later be used to map fresh data. There are numerous supervised learning algorithms available, including Naive Bayes classifiers, Neural Nets, and Support Vector Machines. Unsupervised learning uses data that has not been labeled and trains models without using any pre-defined datasets. Unsupervised learning can be viewed as a powerful tool for examining available data and looking for patterns and trends. Unsupervised learning employs a variety of techniques, including hierarchical clustering and K-means clustering.

2. BACKGROUND

2.1 Data Mining

Data Mining means how to extract knowledge from data, it can be defined as Knowledge mining from data, knowledge extraction, pattern analysis, or data archaeology. Many people refer to data mining as a Knowledge Discovery from Data (KDD) while others consider data mining as a step in the process of knowledge discovery from data.

2.2 Knowledge Discovery from Database (KDD)

The Knowledge Discovery from Database (KDD) process consists of several iterative steps: -

1. Data Cleaning: to remove noise and inconsistent data.
2. Data integration: where multiple data sources can be combined.
3. Data selection: the relevant data were retrieved from database.
4. Data transformation: The process of transforming data through summary or aggregate procedures into forms suitable for mining.
5. Data Mining: essential process where intelligent methods are applied to extract data patterns.
6. Pattern evaluation: to identify the patterns representing knowledge.
7. Knowledge presentation: where visualization techniques are used to present the mined knowledge.

Therefor data mining is the process of discovering interesting patterns and knowledge from large amount of data. The data sources can include databases, data warehouses, web, or information repositories [11].

2.3 Educational Data Mining (EDM)

To improve students' comprehension of the learning process and to concentrate on discovering, extracting, and assessing factors relevant to students' learning processes, data mining techniques can be applied in the educational field. Educational Data mining is concerned with creating strategies that collect knowledge from information originating in educational settings. Many techniques, including Decision Trees, Neural Networks, Naive Bayes, K-Nearest Neighbor, and many others, are used in educational data mining. Many activities that can be utilized to examine student performance are offered by data mining. The classification task will be used to evaluate student's performance. Finding a model that defines and separates different data classes or concepts is the process of classification. The model can be derived based on the analysis of a set of training data.

2.4 Classification

Classification is a form of data analysis that extracts models describing important data classes called classifiers used to predict categorical class labels, for example to build a classification model to categorize bank loan applications as either safe or risky. This analysis can help provide us with a better understanding of the data at large. Many classification methods have been proposed in machine learning, pattern recognition, and statistics. Several industries, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnostics, use classification.

2.5 Clustering

Clustering is the method of converting a group of abstract objects into classes of similar objects, clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters. A cluster is a subset of similar objects such that the distance between any of the two objects inside the cluster is less than the distance between any object that is not located inside the cluster, data objects of a cluster can be considered as a one group. Information set will be partitioned into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups.

Clustering analysis is frequently utilized in a variety of fields, including data analysis, market research, pattern identification, and image processing.

2.6 Association Rules

In many different types of databases, association rules if-then statements help to illustrate the likelihood of links between data items in big data sets. In order to find sales correlations in transactional data or in medical data sets, association rule mining is frequently utilized. Association rules are employed in data science to discover correlations and co-occurrences between data sets. They are ideally used to explain patterns in data from seemingly independent information repositories, such as relational databases and transactional databases. Association rule mining or mining associations are other names for the practice of employing association rules. Association rule mining involves the use of machine learning models to analyze data for patterns, or

co-occurrences in a database. It identifies frequent if-then associations that are the association rules. The antecedent (if) and the consequent (then) are the two components of an association rule. A data point located within the data is an antecedent. An item discovered beside the antecedent is known as a consequent. Association rules are helpful in data mining for studying and forecasting consumer behavior. Customer analytics, market basket analysis, product clustering, catalogue design, and retail layout all benefit from their use.

3. DATA MINING ALGORITHMS

There are a number of data mining algorithms that are implemented in other research papers. Different datasets will produce different results based on the algorithms used. In this study, the proposed approach will be testing some algorithms based on decision trees, rule-based classification, Bayes theorem, and neural networks. The Research aim is to find the best algorithm that gets the high accuracy rate.

3.1 Decision Tree

Decision tree is one of the easier data structures to understand data mining. To create the decision tree, rules from the training dataset are first extracted, and the testing dataset is then classified using the decision tree algorithms. A decision tree is necessarily a tree with an arbitrary degree that classifies instances. Decision tree is a powerful tool for classification and prediction but requires extensive computation. Creating the tree based on the training set takes time although making decisions once the tree is made is not time consuming. Classification tree algorithms may be divided into two groups: one whose result is a binary tree and the other non-binary tree. In decision trees, the leaf node represents the complete classification of a given instance of the attribute and the decision node specifies the test that is conducted to produce the leaf node. With a decision tree, the sub tree that is created after any node is necessarily the outcome of the test that was conducted. A decision tree is used to categorize a certain instance from the tree's root up to the leaf node that shows the instance's result. A major issue in using decision tree is to find out how deep the tree should grow and when it should stop. Usually if all the attributes are different and lead to the same outcome, the decision tree might not be the most effective for decision making and the size of the tree will be large. There are a number of algorithms that are based on decision trees. Some of the most common and effective types of algorithms based on decision trees are C4.5 algorithm and Classification and Regression Tree (CART) Algorithm. Weka is based on the C4.5 learning algorithm, and the C4.5 is a modified version of the basic ID3 algorithm.

3.1.1 C4.5 Algorithm

C4.5 is based on Hunt's algorithm. To create a decision tree, C4.5 manages categorical and continuous attributes. Based on the threshold, C4.5 divides the attribute values into two groups, treating all values above the threshold as one child and the remainder values as another. Gain Ratio is used by C4.5 to construct a decision tree as an attribute selection metric. It eliminates information gain bias when there are numerous result values for an attribute. The multiway split decision tree includes the C4.5 algorithms. C4.5 yields a binary split if the selected variable is numerical, but if there are other non-numerical values representing the attributes it will results in a categorical split. That is, the node will be split into N nodes where N is the number of categories for that attribute. The J4.8 decision tree algorithm in WEKA is based on the C4.5 decision tree algorithm.

3.2 Random Forest Algorithm

An approach for machine learning that uses supervised learning is called Random Forest. It can be used for both Classification and Regression problems in Machine Learning (ML). It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. The Random Forest classifier employs many decision trees on distinct subsets of the input dataset, averaging the outcomes to enhance the dataset's anticipated accuracy. Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and over fitting are prevented by the larger number of trees in the forest.

3.3 Naïve Bayes Algorithm

Based on the Bayes theorem, the Nave Bayes algorithm is a supervised learning technique for classification problems. It is mostly used for categorizing texts using a sizable training set. One of the most straightforward and efficient classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object, it is called Bayes because it depends on the principle of Bayes' theorem. The Bayes' theorem, often known as Bayes' Rule or Bayes' law, is used to calculate the likelihood of a hypothesis given some prior information. It depends on the conditional probability. The Naïve Bayes classification reads a set of examples from the training set and uses the Bayes theorem to estimate the probabilities of all classifications. For each instance, the classification with the highest probability is chosen as the prediction class.

3.4 Multi-Layer Perceptron Algorithm

An input layer, a hidden layer, and an output layer are the minimum number of layers of nodes that make up a multilayer perceptron (MLP), a fully connected class of feedforward ANN. Every node uses a nonlinear activation function; the input nodes are the only exception. Backpropagation is a supervised learning method that is used by MLP during training. The several layers and non-linear activation of MLP set it apart from a linear perceptron. It can discriminate between data that cannot be separated linearly. The outputs of certain neurons become the inputs of other neurons when neurons are combined to form a neural network. One input layer with one neuron (or node) for each input, one output layer with one node for each output, and an arbitrary number of hidden layers with an arbitrary number of nodes in each hidden layer are the constituents of a multi-layer perceptron.

3.5 OneR Algorithm

OneR, short for One Rule Learner, is a classification algorithm that creates one rule for each predictor in the data before choosing the rule with the smallest overall error to be its one rule. It is used in the sequential learning algorithm to learn the rules. It produces a single rule that at least partially addresses the examples. Its power is in its capacity to establish relationships between the provided attributes, hence encompassing a wider hypothesis space. The greedy searching paradigm used by the Learn-One-Rule algorithm allows it to find the rules with great accuracy but with a very limited coverage. It organizes all the uplifting examples for a specific situation. It provides one rule that addresses a few examples. The broadest rule precondition is used as a starting point, and the variable that best boosts performance as tested against the training instances is then eagerly added. Sequential Learning Algorithm uses this algorithm.

4. RELATED WORK

[Mustafa Yagci, 2022][36] conducted research (Educational Data Mining: Prediction of students' Academic Performance using Machine Learning Algorithms) the researcher collected a data of students' university called Kirsehir Ahi Evran university of 1854 record for each student taking Turkish language-1 course. Each student's record contains the midterm exam grades, final exam grades, faculty name, and department name. The researcher used Orange machine learning software to analyze the students' data to employ Random Forest, Neural Network, Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbor models, the researcher purpose is to make predictive and descriptive models. The performance of the models was evaluated through confusion matrix, classification accuracy, precision, recall, f-score, Receiver Operating Characteristics (ROC), and Area Under the Curve (AUC). The highest accuracy rate was obtained by the random forest algorithm with 74.6%, the other classifiers neural network, support vector machine, logistic regression, naïve bayes, and k-nearest neighbor received 74.6%, 73.5%, 71.7%, 69.9% accuracy percentage respectively.

[Lonia Masangu, Ashwini Jadhav, Ritesh Ajoodha, 2021][37] conducted research about (Predicting Student Academic Performance using Data Mining Techniques). The researchers obtained the dataset from Kaggle website consisted of 480 student's records for undergraduate students in school from grade 1 to grade 12. The students were categorized into Low, Medium, and High according to their grades. The students' data contains academic information such as grades, students' educational stage, demographic information such as nationality, gender, and behavioral information such as number of raised hand, number of accessed forums for each student. To analyze the data, the researchers applied 5 data mining algorithms (Support Vector Machine, Decision Tree, Perceptron, Logistic Regression, Random Forest). Preprocessing activity applied to the data to change the class label "Student grades" from categorical (low, medium, high) to numeric (-1,0,1) respectively to preserve the distance between the categorical values and scale the numeric field so they can be meaningful when they compared together. The obtained results showed that the student absence days' field influenced the student academic performance rather than the student's grades. Support vector machine algorithm outperformed other classifiers with accuracy rate 70.8%. The other classifiers Decision Tree, Perceptron, Logistic regression, random forest achieved accuracy rate with 46.8%, 64.5%, 67.7%, 69.7% respectively.

[Khaledun Nahar, Boishakhe Islam Shova, Tahmina Ria, Humayara Binte Rashid, A.H.M.Saiful Islam, 2021] [5] Presented a comparative study of data mining techniques (Mining Educational Data to Predict Students Performance). The researchers collected educational data from CSE department of Notre dame university Bangladesh of 80 students from mark sheets and survey. The researchers created 2 datasets to focus on different angles, the first one classifies and predicts the category of students (good, bad, medium) on artificial intelligence course based on their pre-requisite course performance, the second dataset classifies and predict the final grade of any random subject. The researchers used WEKA open-source software to apply 6 algorithms (Decision Tree J48, Naïve Bayes, PART, Bagging, Boosting, Random Forest). Two models have been created, the first one was based on the decision tree which was constructed by analyzing the first dataset named as AI Prerequisite dataset and gives 64.3% accuracy on the test dataset, and the second model was based on the Naive Bayes created according to the second dataset named Theory Performance and gives 75% accuracy rate on the test dataset.

[Anupam Khan, Soumya K. Ghosh, 2021] [2] conducted a review on educational data mining studies (Student Performance Analysis and Prediction in classroom learning: A review of educational data mining studies). The researchers presented a review of 140 studies in this area, they focused on studies related to classroom-based education. They presented meta-analysis of performance influencing factors, aim, and the time of prediction. The researchers observed that the factors that influenced the students' performance were the grades of earlier attended courses, quizzes, midterm examination, student's behavior, quality, and family background. Teaching quality and domain knowledge also influences student performance. Most studies use Regression and classification for establishing the impact of various factors. Some studies tried to predict students' performance before course commencement. However, they were less efficient compared to the prediction studies during the tenure of the course. Internal assessment and behavior seem to be the most effective predictors of student performance during the tenure of the course, but these predictors were not available before course commencement. The researchers observed that most of the studies focused on student success prediction rather than the prediction of the final score.

[Ferda Unal, 2020] [3] presented research on (Data Mining for Student Performance Prediction in Education), The researcher used two public datasets from Portuguese school, each dataset consists of 33 attributes about student's grades, social demographic, and school related features; the first dataset had information regarding the students' performance regarding Portuguese Language lesson and the other one was about mathematics lesson. Pre-processing activities were applied on the datasets before applying the classifiers; The first pre-processing activity was dividing the data by two different grade categorizations: Five grade categorization and Binary grade categorization. Weka open software was used to employ three classification algorithms (Decision Tree J48, Random Forest, Naïve Bayes). The best performance for the five levels grading on language dataset was achieved by Random Forest algorithm by 73.50% while in the binary grading this percentage had increased to 93.07% with the same random forest algorithm. On the mathematics dataset, the five levels categorization, the best performance was achieved by the Decision Tree algorithm (J48) by 73.42%, while at the binary grading Random Forest achieved the highest percentage with 91.39%. Then another pre-processing activity was applied to enhance the performance of the classifiers which had been called Wrapper method for feature subset selection to find the optimal subsets of features. The Wrapper subset method had a recursive structure, the process started with selecting a subset and inducing algorithm on that subset, then the evaluation was made according to the success of the model. There were two options in this assessment: the first one returned to the top to select a new subset; the second option used the currently selected subset. The wrapper subset attribute selection method was applied on the five-level grade categorization on both language and mathematics datasets among the three algorithms (Decision Tree J48, Random Forest. Naïve Bayes). The same procedure applied on the binary level grade categorization. Wrapper subset method was used to increase the accuracy rate. One of the important steps to create a good model is the attribute selection; this operation can be done in two ways: first to select the relevant attributes and second to remove the redundant or irrelevant attributes. Attribute selection was used to create a simple model. On language and mathematics datasets. The accuracy rates had changed positively in all trails using wrapper subset attribute selection method. Wrapper feature subset selection method was used to improve the classification performance. Preprocessing operations was applied on the dataset. Categorizing the final grade field into five and two groups increased the percentage of accurate estimates in the classification. The wrapper attribute selection method in all algorithms has led to a noticeable increase in accuracy rate. Overall, better accuracy rates were achieved with the binary class method for both mathematics and language datasets.

[Lili Dwi Yulianto, Agung Triayudi, Ira Diana Sholihati, 2020] [6] presented a study about (Implementation Educational Data Mining for Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5). The researchers collected data of students from the National University of Jakarta consists of 588 records collected through questionnaire and literature study, pre-processing activities were done to eliminate the unnecessary attributes from the dataset. The researchers used RapidMiner Studio 9.2.000 and PHP programming language to apply decision tree C4.5 and K-Nearest Neighbor (KNN) algorithms to predict the students' performance and specify the most significant attributes. The highest gain ratio was for the four attributes (GPA, study hours, scholarship, and graduation status). Cross validation and Confusion matrix were used for testing and evaluation. The highest accuracy rate was obtained by the K-Nearest Neighbor (KNN) algorithm with 59.32% while the decision tree C4.5 algorithm gets 54.80% accuracy rate.

[Raza Hasan, Sellappan Palaniappan, Salman Mahmood, Ali Abbas, Kamal Uddin Sarker, Mian Usman Sattar, 2020] [7] conducted research about (Predicting Student Performance in Higher Educational Institutions using Video Learning Analytics and Data Mining Techniques). The researchers obtained data from Higher Education Institute (HEI) in Oman academic year 2019 for 2 subjects for 772 students, 19 attributes were selected. The researchers get the students data from Student Information System (SIS), Learning Management System (LMS), and Video Streaming Server (VSS). The researchers used mobile application called eDify to present lectures as videos for students before class commencement. This approach of education is called flipped classroom, attributes related to the video streaming were considered in the test data for the model built such as video played, paused, liked, and video segmentation. The researchers had used Orange data mining tool to apply 8 algorithms (K-Nearest Neighbor KNN, Neural Network, Logistic Regression, Decision Tree, CN2 Inducer, Random Forest, Naïve Bayes, and Support Vector Machine SVM). A supervised data classification technique was used to determine the best prediction model that fit the requirements. The same set of classification algorithms, performance metrics, and the 10-fold cross-validation methods were used. Three feature selection methods were used: information gain, information gain ratio and Gini decrease. Nine attributes were selected from ranking and scoring techniques. First the performance of the most widely used classification algorithms was compared using the complete dataset. Second data transformation and feature selection techniques were applied to the dataset to determine the impact on the performance of the classification algorithms. Lastly, they reduced the features and determined the appropriate features that can be used in order to predict the students' performance. Four performance metrics (accuracy, sensitivity, specificity and F-Measure) were evaluated in order to compare the performance of the classification models. The genetic algorithm was used to further reduce the features; attributes were reduced from 9 to 6 attributes. Principle Component Analysis (PCA) was undertaken to reduce the number of attributes from 19 to 8 attributes with variance 95.6% but PCA resulted in one fewer component. The highest accuracy result was obtained through Information Gain Ratio Feature Selection and Equal Width Transformation by Random Forest algorithm by 88.3% accuracy, 97.5% sensitivity, 89.5% specificity, and 93.3% F-Measure. CN2 Inducer has the second-high accuracy with 87.4 %. Out of 45 rules, 23 were selected through it.

[Y. K. Salal, S. M. Abdullaev, Mukesh Kumar, 2019] [10] conducted a research based on (Educational Data Mining: Student Performance Prediction in Academic). The research was made about two datasets from two secondary schools from Portuguese. The dataset was consisted of 33 attributes includes academic grades, demographic attributes, social attributes, and school related attributes. Reports and questionnaires were used for collecting data from students. There were different algorithms applied on the datasets through Weka open-source software such as Naïve Bayes, Decision Tree (J48), Random Forest, Random Tree, REPTree, JRip, OneR, Simple Logistic and ZeroR. Tested using 10-fold cross validation with different parameter selection experimented first time with the all 33 attributes and the second time with only 8 attributes. The attributes were evaluated through Correlation, Gain Ratio, and Info Gain search methods. The REPTree and OneR algorithms performed well with accuracy 76.7334 %. The overall results of the algorithms OneR, REP Tree, and Decision Tree (J48) were more than 76% accuracy for predicting student results and they performed equally well. Other classification algorithms such as Simple Logistic, JRip, and Naïve Bayes get more than 73% accuracy, while Random Tree classification algorithms get the overall accuracy less than 70%. The most significant attributes were G2 (second term grade), G1 (First Term Grade), School name, and study time hours for the prediction of the students' performance.

[Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, Norafefah Mohd Sobri, 2019] [8] presented a study about (Supervised Data Mining Approach for Predicting Student Performance). The researchers collected data of undergraduate students of Bachelor of Science in Statistics program from faculty of Computer and Mathematical Sciences at two Universities in Malaysia consisted of 631 records for three academic years from 2013 to 2016, each student has attributes of id, CGPA (cumulative grade point average) coded into 1=Excellent or 2= Not-Excellent, and grades for 11 subjects. The researchers followed the CRISP-DM methodology (Cross Industry Standard Process for Data Mining). It is a cyclic approach consists of 6 stages (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment), then pre-processing activities were applied to the data. The researchers used RapidMiner 8.3 software application to apply 5 algorithms (K-Nearest Neighbor KNN, Naïve Bayes, Decision Tree information gain, Decision Tree Gini precision, Logistic Regression). Receiver Operating Characteristics (ROC) curve was used for the evaluation of classification algorithms. The results indicated that Naïve Bayes Classifier outperformed the other classifiers with 89.26% accuracy rate while logistic regression achieved 85.28% accuracy. All the supervised classifiers were performing above 80% which shows that the error rate is low and the predictions are reliable.

[Jabeen Sultana, M. Usha Rani, M. A. H. Farquad, 2019] [4] presented a study about (Student's Performance Prediction using Deep Learning and Data Mining Methods). The researchers collected data from a Saudi university database from learning management system called D2L consisted of 1100 student records with 11 attribute each. Pre-processing filtering activities were done to the data then it fed to the WEKA software to apply eight different algorithms Multiple Layer Perceptron (MLP), Multi-Class Classifier, Support Vector Machine (SVM), Naïve Bayes, Instance Based k (IBK), Lazy Locally Weighted Learning (LWL), Random Forest, Decision Tree (J48). Student's records will be classified into three class labels (Low, Medium, and High). Classifiers were evaluated through Specificity, Sensitivity, Kappa-statistics, and Receiver Operating Characteristics (ROC) curve. The results indicated that the optimal results were obtained by Multiple Layer Perceptron (MLP), Multi-Class Classifier, Random Forest, and Decision Tree J48 with accuracies 99.45%, 99.81%, 100%, and 100% respectively.

[Amjed Abu Saa, Mostafa Al-Emran, Khaled Shaalan, 2019] [1] presented a systematic literature review about (Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques). The researchers reviewed and analyzed 36 research articles from 2009 to 2018 by applying Systematic Literature Review (SLR) approach. The aim of this SLR is to identify the most commonly studied factors affects students' performance and the most common data mining techniques applied to identify these factors. The researchers found that the most frequent data collection techniques used were the e-learning system logs and student information system data. The most significant factors affect the students' performance in higher education were the students' previous grades, students' e-learning activity, students' demographics, and students' social information. The most commonly used data mining algorithms were 9 algorithms (Naïve Bayes Classifiers, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor (KNN), ID3 Decision Tree, C4.5 Decision Tree, Decision Tree (DT), Multi-Layer Perceptron Neural Network (MLP), Neural Network (NN)). Classification approach was the most commonly used approach among the analyzed studies; only four research papers have used clustering with classification technique to find out how many groups of students were available in the dataset and extract specific features of each group.

[Y. K. Saheed, T. O. Oladele, A. O. Akanni, W. M. Ibrahim, 2018] [9] presented research about (Student Performance Prediction based on Data Mining Classification Techniques). The researchers worked on a dataset from private university in northern part of Nigeria faculty of Computer Science with 234 records. Pre-processing activities and attributes selection processes were applied and ended up with 13 attributes for each record, Weka software was used for the analysis through applying decision tree algorithms such as ID3, C4.5, and CART algorithms. The results obtained showed that the C4.5 classification algorithm outperformed other classifiers with accuracy prediction 98.3 %. The most significant attributes were the courses applied for admission like Bachelor degree, socio-demographic factors like age, and parental factors like mother, father occupation.

[Hilal Almarabeh, 2017] [12] conducted research about (Analysis of Students' Performance by Using Different Data Mining Classifiers). The study used five classifiers (Naïve Bayes, Bayesian Network, ID3, J48, Neural Network). The researcher gets dataset from students' college database consisted of 225 record 10 attributes each record. Weka open-source software was used for the analysis. Different error measurement methods were used for the evaluation such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). The Naïve Bayes algorithm outperformed other classifiers with 92.0 % accuracy rate.

[Mukesh Kumar, A.J. Singh, 2017] [13] presented research about (Evaluation of Data Mining Techniques for Predicting Students' Performance). The dataset was about post-graduate students about their graduation marks, financial condition of the family, father and mother education (personal attributes, family attributes, academic attributes, institutional attributes) for 412 students' record. The dataset was fed into Weka open-source software to apply different algorithms (Decision Tree J48, Naïve Bayes, random Forest, PART, and Bayes Network). Several evaluation methods were used such as Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error, Receiver Operating Characteristics (ROC), Area Under Curve (AUC). The evaluation was done first through Training dataset mode, second with 10-fold cross-validation mode, then with percentage split mode with 74% percentage split. Random forest gave the best result with 61.40% for the correctly classified instances.

[Annisa Uswatun Khasanah, Harwati, 2017][14] presented research about (A Comparative Study to Predict Students' Performance using Educational Data Mining Techniques). The dataset was collected from faculty of Industrial Engineering Islam Indonesia University's Information System. The dataset consisted of 104 students' records, 12 attributes each with class label drop out or not; there were 13 records classified as drop out and 91 records classified as not dropped out. Pre-processing activities were applied to the data, and then feature selection methods were used to select the most significant attributes such as correlation-based attribute evaluation, gain ratio attribute evaluation, information gain attribute evaluation, relief attribute evaluation, and symmetrical uncertainty attribute. Weka open-source software was used to apply decision tree algorithms and Bayesian network algorithm. The most significant attributes were first semester GPA, first semester attendance, senior high school department, gender, father occupation, mother occupation, and mother education. Bayesian network gets higher accuracy rate than the decision tree algorithm with 98.08% and 94.23% respectively.

[Adelaja Oluwaseun Adebayo, Mani Shanker Chaubey, 2019] [15] presented a study about (Data Mining Classification Techniques on the Analysis of Students' Performance). The researchers used KNIME analytics platform which is open-source software for creating data science applications, continuously integrating new developments from machine learning and data mining. In KNIME every object in the dataset is assigned to a pre-defined class label. The building of models with good generalization capability is a key objective of the learning algorithm, for instance models that accurately predict the class labels of previously unknown records. The built model in the KNIME is used in:

- Classification of unknown objects performed based on the constructed model.
- Result class label will be compared with the class label of test sample.
- Calculation of the percentage of test sample and accuracy of model should be compared with training sample.
- There are always differences between the test sample data and training sample data.

The constructed model read school marks data from excel sheet to construct a decision tree with four levels. Histograms for each tree level were generated and also pie-charts for each decision tree.

[Raheela Asif, Saman Hina, Saba Izhar Haque, 2017] [38] presented study about (Predicting Student Academic Performance using Data Mining Methods). The researchers collected data of undergraduate students at faculty of Civil Engineering at university in Pakistan for 214 student's records. Each student had attributes of High School Certificate, first, and second years' course's grades.

The researchers used Rapid Miner tool to analyze the data through Decision Tree with Gini Index, Decision Tree with Information Gain, Decision Tree with Accuracy, Neural Networks, Naïve Bayes, Random Forest with Gini Index, Random Forest with Information Gain, and Random Forest with Accuracy. The researchers applied different feature selection techniques available in Rapid Miner tool like Recursive Feature Elimination (RFE), it has four criterions to weight attributes: weight by Gini Index, weight by Information Gain, weight by chi-square, weight by rule induction to select subset of the attributes in order to increase the accuracy rate. The results showed that Decision Tree with Gini Index, Decision Tree with Information Gain, Decision Tree with Accuracy, Neural Networks, Naïve Bayes, Random Forest with Gini Index, Random Forest with Information Gain, and Random Forest with Accuracy get accuracy rate with 74.78%, 73.91%, 55.65%, 68.70%, 74.78%, 67.83%, 67.83%, 64.35% respectively.

[Nawal Ali Yassein, Rasha Gaffar M Helali, Somia B Mohamad, 2017] [39] conducted research about (Predicting Student Academic Performance in KSA using Data Mining Techniques). The researchers collected the data from faculty of computer science and business administration at Najran university for 150 students' records and 108 courses records. Each student had 7 attributes (attendance, lab work, mid-term grades, assignments, exam marks, education type, and success rate). If the success rate is greater than 65% it is classified as High, if it is lower than 65% it is considered Low. And each course has 12 attributes (course id, credit hours, practical work, assignments, number of assignments, midterm exams, number of midterm exams, number of final exam questions, education type, course description, study field, success rate). The researchers used Clementine data mining toolkit to apply feature selection algorithm and clustering algorithm through Statistical Package for Social Sciences (SPSS) to identify the attributes that affect course success rate. The model developed employed both classification and clustering techniques to identify features affecting student's performance in selected courses. Records were randomly split into two sets, a training set and a testing set. The training set was used to create the mining model. The testing set was used to check the model accuracy. Training data represented 40% of total data/records. Courses records fed firstly to two steps clustering algorithm and feature selection algorithm. Student records was ranked as important, unimportant, and neutral. Clustering was done on the dataset and the algorithm classified data into two clusters that confirmed the suitability of the selected attributes for prediction purpose. Then for supervised classification; data were labeled according to success rate either High or Low. Data was further fed to C4.5 algorithm to find out which of the selected attributes could be used as a predictor for the success rate. The previous steps were repeated for student's success rate prediction. Prediction accuracy was estimated by the used mining algorithm equal to 100% for both student success rate prediction and course prediction. The model gets precision and recall equal to 1. The implemented model showed a strong relation between practical work attribute with its success rate and student attendance with its success rate.

5. EDUCATIONAL DATA MINING CHALLENGES

Despite the many advantages offered by predicting students' performance using data mining techniques, there are still some associated challenges.

- There are few available datasets to be used in the prediction of students' academic performance in the university bachelor degree stage.

- There are rarely datasets for the undergraduate students in computer science field to be used in the prediction of students' academic performance.
- The attributes that are considered as dominant factors for student's performance prediction are unknown.
- Prediction made by data mining algorithms is not completely accurate.
- The way attributes of the dataset used to evaluate using different ranker methods such as finding Correlation between attributes is difficult to predict.

6. CONCLUSION & FUTURE WORK

Prediction of students' performance will help the universities guide students, who are likely to have bad results, in advance to do better or to recommend more suitable courses for them. In this study various data mining techniques are described for the students' performance detection that had been proposed the past few years. The solution is to develop an efficient approach for prediction of student performance in order to prevent the problems that were discussed previously. Hybrid approaches are suggested to be used in providing better results and overcome the drawbacks. As a result, this may yield more accurate results [35].

7. REFERENCES

[1] Amjed Abu Saa, Mostafa Al-Emran, Khaled Shaalan, "Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques", Springer Nature, pp. 567-598, April 2019.

[2] Anupam Khan, Soumya K. Ghosh, "Student Performance Analysis and Prediction in Classroom Learning: A Review of Educational Data Mining Studies", Springer Nature, pp. 205-240, January 2021.

[3] Ferda Unal, "Data Mining for Student Performance Prediction in Education", The Graduate School of Natural and Applied Sciences, Dokuz Eylul University, Izmir, Turkey, License IntechOpen, pp. 1-11, 2020.

[4] Jabeen Sultana, M. Usha Rani, M.A.H. Farquad, "Students' Performance Prediction using Deep Learning and Data Mining Methods", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Vol.8, pp. 1018-1021, June 2019.

[5] Khaledun Nahar, Boishakhe Islam Shova, Tahmina Ria, Humayara Binte Rashid, A.H.M.Saiful Islam, "Mining Educational Data to Predict Students Performance A Comparative Study of Data Mining Techniques", Springer Nature, pp. 6051-6067, May 2021.

[6] Lili Dwi Yulianto, Agung Triayudi, Ira Diana Sholihati, "Implementation Educational Data Mining for Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5", Journal Mantik, ISSN: 2685-4236, Vol.4, pp. 441-451, May 2020.

[7] Raza Hasan, Sellappan Palaniappan, Salman Mahmood, Ali Abbas, Kamal Uddin Sarker, Mian Usman Sattar, "Predicting Student Performance in Higher Educational Institutions using Video Learning Analytics and Data Mining Techniques", Applied Sciences, pp. 1-20, June 2020.

[8] Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, Norafefah Mohd Sobri, "Supervised Data Mining Approach for Predicting Student Performance", Indonesian Journal of Electrical Engineering and Computer Science, ISSN: 2502-4752, Vol.16, pp. 1584-1592, December 2019.

[9] Y. K. Saheed, T. O. Oladele, A. O. Akanni, W. M. Ibrahim, "Student Performance Prediction based on Data Mining Classification Techniques", Nigerian Journal of Technology (NIJOTECH), ISSN:0331-8443, Vol.37, pp. 1087-1091, October 2018.

[10] Y. K. Salal, S. M. Abdullaev, Mukesh Kumar, "Educational Data Mining: Student Performance Prediction in Academic", International journal of Engineering and Advanced Technologies (IJEAT), ISSN: 2249-8958, Vol.8, pp. 54-59, April 2019.

[11] Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING Concepts and Techniques" Book Third Edition.

[12] Hilal Almarabeh, "Analysis of Students' Performance by Using Different Data Mining Classifiers", I.J. Modern Education and Computer Science IJMECS, pp. 9-15, August 2017.

[13] Mukesh Kumar, A.J. Singh, "Evaluation of Data Mining Techniques for Predicting Students' Performance", I.J. Modern Education and Computer Science IJMECS, pp. 25-31, August 2017.

[14] Annisa Uswatun Khasanah, Harwati, "A Comparison Study to Forecast Students' Performance Using Educational Data Mining Methods", Materials Science and Engineering IOP Conference Series: Materials Science and Engineering, DOI: 10.1088/1757-899X/215/1/012036, pp. 1–7, 2017.

[15] Adelaja Oluwaseun Adebayo, Mani Shanker Chaubey, "Data Mining Classification Techniques on the Analysis of Students' Performance", Global Scientific Journals, ISSN: 2320-9186, Vol.7, Issue 4, pp. 79-95, April 2019.

[16] saumyasaxena2730, ayushgangwar, tanushree7252 riddhijain1826, avinashrayz28, bhandarimayvp1f, "Basic Concept of Classification Data Mining" last updated: 06 May 2023, Geeks for Geeks website, accessed: July 2023, <https://www.geeksforgeeks.org/>.

[17] Amrieh, E. A., Hamtini, T., & Aljarah, "Students' Academic Performance Dataset", 08 November 2016, Kaggle website, accessed: November 2022, < https://www.kaggle.com/>.

[18] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

[19] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

[20] Arun George Eapen, "Application of Data mining in Medical Applications", 2004, collectionscanada.gc.ca website, accessed: November 2022, <https://www.collectionscanada.gc.ca>.

[21] Robin Norberg, “Master of Science in Engineering Technology Computer Science and Engineering”, 2013, DiVA portal website, accessed: November 2022, <https://www.divaportal.org>.

[22] “Weka Tutorial”, Tutorials point website, accessed: October 2022, <https://www.tutorialspoint.com/weka/index.htm>.

[23] “Weka 3: Machine Learning Software in Java”, The University of Waikato website, accessed: October 2022, <https://www.cs.waikato.ac.nz/ml/weka/>.

[24] “Weka Data Mining”, Javatpoint website, accessed: October 2022, <https://www.javatpoint.com/weka-data-mining>.

[25] “What is Text Mining, Healthcare Natural Language Processing, and LLMs”, 27 November 2023, Linguamatics website, accessed: November 2023, <https://www.linguamatics.com>.

[26] Sara Brown, “Machine learning explained”, 21 April 2021, MIT Sloan website, accessed: November 2022, <https://mitsloan.mit.edu/>.

[27] “Machine Learning”, last update: 14 January 2024, Wikipedia website, accessed: 17 January 2024, <https://en.wikipedia.org/wiki/Machine_learning>.

[28] “Clustering in Data Mining”, Javatpoint website, accessed: November 2022, <https://www.javatpoint.com/data-mining-cluster-analysis>.

[29] Ben Lutkevich, “association rules”, last updated: June 2023, Tech Target website, accessed: July 2023, <https://www.techtarget.com>.

[30] Carine Bou Rjeily, Georges Badr, Amir Hajjam El Hassani, Emmanuel Andres, “Sequential Mining Classification”, 23 October 2017, IEEE Xplore website, accessed: July 2023,<https://ieeexplore.ieee.org/document/8079747>

[31] Craig Stedman, Adam Hughes, “Data Mining”, last updated: September 2021, Tech Target website, accessed: July 2023, <https://www.techtarget.com>.

[32] “What is Data Mining?”, IBM website, accessed: July 2022, <https://www.ibm.com/topics/data-mining>.

[33] “Educational Data Mining”, Educational Data Mining website, accessed: August 2022, <https://educationaldatamining.org/>.

[34] Nikita Jain, Vishal Srivastava, “Data Mining Techniques: A Survey Paper”, International Journal of Research in Engineering and Technology IJRET, eISSN: 2319-1163, Volume: 02, pp. 116-119, November 2013.

[35] Shikha Agrawal, Jitendra Agrawal, “Survey on Anomaly Detection using Data Mining Techniques”, Procedia Computer Science, pp. 708-713, December 2015.

[36] Mustafa Yagci, “Educational Data Mining: Prediction of students’ academic performance using machine learning algorithms”, Springer Open, Smart Learning Environment, pp. 9-11, March 2022.

[37] Lonia Masangu, Ashwini Jadhav, Ritesh Ajoodha, “Predicting Student Performance using Data Mining Techniques”, Advances in Science, Technology and Engineering Systems Journal ASTES Journal, ISSN: 2415-6698, Vol. 6, pp. 153-163, 2021.

[38] Raheela Asif, Saman Hina, Saba Izhar Haque, “Predicting Student Performance using Data Mining Methods”, International Journal of Computer Science and Network Security IJCSNS, Vol.17, pp. 187-191, May 2017.

[39] Nawal Ali Yassein, Rasha Gaffar M Helali, Somia B Mohamad, “Predicting Student Academic Performance in KSA using Data Mining Techniques”, Journal of Information Technology & Software Engineering, ISSN: 2165-7866, Vol. 7, pp. 1-5, 2017.