IJCSMA

Amal Abdulbaqi Maryoosh *et al*, International Journal of Computer Science and Mobile Applications,
Vol.10 Issue. 3, March- 2022, pg. 1-14 **ISSN: 2321-8363**
**Impact Factor: 6.308**
**(An Open Accessible, Fully Refereed and Peer Reviewed Journal)**

# A Review: Data Mining Techniques and Its Applications

**Amal Abdulbaqi Maryoosh[1]; Enas Mohammed Hussein[1]**
[1]Computer Science Department, Faculty of Education, Mustansiriyah University, Baghdad, Iraq

**DOI: 10.47760/ijcsma.2022.v10i03.001**

## Abstract

Data mining is a set of processes by which knowledge is extracted from huge amounts of data. Data mining is used to extract useful patterns and hidden information from this data. Machine learning techniques help in the comprehension of the hidden knowledge in the data. Data mining is considered an important field of research and is used in many different fields such as fraud detection, financial banking, education, healthcare, agriculture, industry, etc. In this paper, we will highlight some fundamentals of data mining and its applications. Also, we will conduct a comparative study among different reviews, combining literary studies that employed data mining techniques in various fields and reviewing the latest developments in this field.

***Keywords**:* Data mining, Classification, Clustering, Association rules, Machine learning.

## 1. Introduction

Data plays a basic role in the development of any organization. It is necessary to extract and analyse this data to discover the information represented by patterns and find the relationships between this data that can be used to make correct predictions. These processes are called data mining (Solanki & Patel, 2015). Data mining is a collection of methods and techniques that allow the analysis of large amounts of data to extract knowledge through the use of algorithms and techniques derived from the fields of machine learning, statistics, and database management systems (Anshu, 2019). There is a worry about missing values, noisy data, sparse data, static data, dynamic data, attractiveness, heterogeneity, significance, data size, algorithm efficiency, and complexity. The data we have is often huge and noisy, which means it is imprecise and it has a complex structure. In this case, the purely statistical technique will not work, so data mining is a solution (Aruna & Butey, 2014). Data mining can be divided into two main parts: descriptive mining and predictive mining, each with different functions and techniques. Data mining techniques are divided into three main groups: statistical techniques, machine learning techniques, and artificial intelligence techniques. Each of these techniques has its own algorithms to create models to get the best solution (Mustafa Abdalrassual Jassim & Abdulwahid, 2021). The process of data mining is a gradual process consisting of several steps. This process is called knowledge discovery. The knowledge discovery process includes the following steps that must be done in order: data cleaning, integration, selection, transformation, mining, pattern evaluation, and knowledge representation (Anshu, 2019). Figure 1 explains the knowledge discovery process.

1. Data cleaning: In this process, noise and conflicting data are eliminated.
2. Data integration: In this process, data is collected from several sources.
3. Data selection: In this process, analysis is used to extract relevant data from the database.

4. Data transformation: In this process, data is standardized using normalization, smoothing, summarization, or gathering to transform it into suitable shapes for the mining process.
5. Data mining: In this step, data patterns are extracted by applying intelligent techniques.
6. With knowledge presentation, the extracted knowledge is visually represented to the user by using visualization techniques to understand the results.

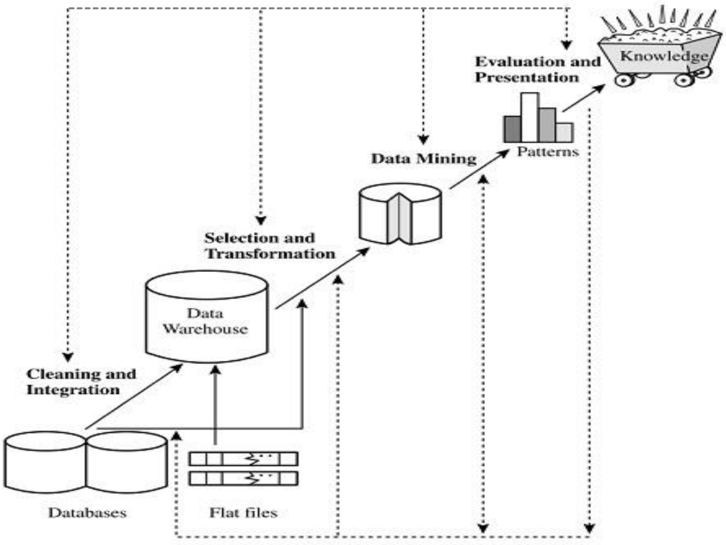

**Figure 1**: the knowledge discovery process (Han & Kamber, 2006)

## 2. Data Mining Techniques

Data mining is the collection of process for gathering relevant information from a vast amount of unstructured data. As mentioned in the Introduction section, the main objective of mining is to create either a descriptive model or a predictive model. In the descriptive model, the data is divided into groups, the total probabilities of the data are distributed, and models are formed to describe the relationships between features. While the predictive model is used to predict future and unknown values (Dogra & TanujWala, 2015; Smita & Sharma, 2014), Predictive and descriptive models contain various algorithms and techniques such as classification, regression, clustering, predication, association rules, etc. All these techniques are used for knowledge discovery.

### 2.1 Classification

Classification is a supervised learning method that is considered the most common data mining technique. It is used to classify each item in a dataset into predefined groups (Pallavi Reddy et al., 2020). In order to extract patterns and rules from data, classification techniques were improved as an essential and significant element in machine learning algorithms that could be used for prediction (Dogra & TanujWala, 2015). There are various techniques that are used in classification, such as Bayesian classification, genetic algorithms, K-Nearest Neighbour, support vector machine, rough set and fuzzy set approach, neural networks, decision trees, logistic and linear regression, and classification based on associations, etc. (Smita & Sharma, 2014; Tamilselvi & Kalaiselvi, 2013). The classification techniques have been used in various applications like identification and prediction in the healthcare industry, credit card fraud detection, spam detection, computer vision, speech recognition, etc. (Anshu, 2019; Ramageri).

## 2.2 Clustering

Clustering is an unsupervised learning method. It does not need a training dataset. Clustering is collecting similar objects together to create a group. In some applications, the clustering is called data segmentation because it segments the large dataset into groups according to its similarity (Han & Kamber, 2006; Solanki & Patel, 2015). Clustering uses variety methods, including partitioning methods (k-means and k-medoids), hierarchical methods (agglomerative and Divisive), density-based methods (Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points to Identify the Clustering Structure (OPTICS), and Clustering Based on Density Distribution Functions (DENCLUE)), grid-based methods (STatistical INformation Grid (STING), and Clustering Using Wavelet Transformation (WaveCluster)), model-based clustering methods (extension of the k-means partitioning algorithm (Expectation-Maximization), conceptual clustering, neural network approach to clustering), high-dimensional data (dimension-growth subspace clustering (CLIQUE), frequent pattern based clustering (pCluster), and dimension-reduction projected clustering (PROCLUS)), and constraint-based clustering analysis (constraints on individual objects, constraints on the selection of clustering parameters, constraints on distance or similarity functions, user-specified constraints on the properties of individual clusters, and semi-supervised clustering based on "partial" supervision)). The clustering techniques can be used in various applications such as data analysis, pattern recognition, market research, image processing, social network analysis, biological and medical data analysis, and outlier detection like identifying fraudulent activity, spam filters, etc. (Han & Kamber, 2006).

## 2.3 Prediction

The numeric prediction, in other words, is called "regression." Regression can be used to represent the relation between one or more independent and dependent variables. In prediction, records are classified according to some predicted future behaviour (Aruna & Butey, 2014; Tamilselvi & Kalaiselvi, 2013). The predictions used numerous other data mining techniques, like some classification techniques (such as support vector machines, backpropagation, and k-nearest-neighbour classifiers) that can be used for prediction (Han & Kamber, 2006; Ramageri).

## 2.4 Association Rules

Association is the most popular data mining technique that describes patterns that tend to occur together in the same transaction. It is used to find the most frequent item sets (Julio Ponce et al., 2009; Sidhant Sethi et al., 2016). The Association Rule is generally used in many fields, such as medical applications, market basket analysis, modern communication networks, etc. Association Rule mining supports many algorithms, especially Apriori, Partitioning, and FP Tree algorithms, processed effectively (Muyeba et al., 2009; P.Thangaraju & D.Nanthini, 2015).

## 3. Survey of Data Mining Applications

Data mining is used in various areas, as it is very beneficial. Some of these fields that have adopted various data mining methods are discussed below.

### 3.1 In Education Field

Lakshmi and A.R.Mohamed Shanavas (Prabha & Shanavas, 2015) made a comparative study of various Educational Data Mining (EDM) algorithms. They compare the performance of classification algorithms Multilayer perceptron, J48, Naive Bayes, ZeroR, and Random Forest. The dataset that is used in this work is taken from students' data who used the MATHSTUTOR (Prabha et al., 2014) e-learning tool. It contains 120 sixth-grade students. The researchers used the WEKA tool to implement the classification algorithms. The results gained in this work are as follows, multi-layer perceptron and J48 have 100% accuracy. J48 takes a very low time of 0.03 seconds compared with 27.19 seconds taken by a multilayer perceptron. The naive bayes algorithm accuracy is 98.33%, and it takes the lowest execution time of 0 seconds. The random forest algorithm has 96.66% accuracy, and the execution time is 0.8 seconds. The ZeroR classifier’s accuracy is very poor at 29%, and the execution time is 0.3.

Raheela Asif and et al. (Asif et al., 2017) used decision trees to evaluate undergraduate students' performance. Their research intends to examine the performance of students pursuing a four-year bachelor's degree in information technology. The dataset that was used in this study was collected from two academic batches using a sample of 210 undergraduate students. The dataset contains variables related to students' grades before their admission to the university (which is used in choosing them for acceptance at university) and the grades for all the courses that are educative in the four years. This study focused upon two aspects of students’ performance. The first aspect is forecasting students' academic performance at the end of a four-year program. The second aspect is studying normal progress and putting it together with forecasted results. They identified two substantial cohorts of students: high and low-achieving students. The study results show that by concentrating on a small number of courses that are predictors of good or poor performance, the outcomes can be obtained, low-achieving students can receive timely warning and help, while high-achieving students can receive advice and opportunity.

Mabel Christina (Christina, 2018) proposed a model that can be used to predict future learning results for students. The dataset was gathered for approximately 350 understudies from the 4-stage B.E. CSE. In this method, a poll shape was used to collect genuine information from the understudies that depicted the relationship between their learning behaviors and their scholastic performance. Understudy statistics points of interest, school subtle aspects, attendance, CGPA, and final review in the previous semester are the factors used in the poll to make a decision regarding the learning and scholastic behavior of understudies. This information was captured in just three weeks. Information from about 300 understudies was acquired. Around 250 datasets were used as preparation datasets, while 50 datasets were used as test information to plan understudy display. Many classification algorithms were used, like MLP, Naive Bayes, SMO, REP tree, Decision Table, and J48, and they were implemented using the WEKA tool. The result demonstrated that the J48 classifier achieved 97%, which is higher accuracy than other classification algorithms. It is used to predict test informative collections for future results, such as best, good, average, or poor.

Manjula V. and M. Srinath (Manjula V & srinath, 2020) used Adaptive Artificial Neural Network (AANN) to identify the reasons for students' weak performance through the use of psychology and sociology the researcher collected 31 attributes for 5000 person. The accuracy gained from implementing AANN is 95%, and this result is the best by comparing it with MLP, Naive Bayes, SMO, REP tree, and J48.

Sarah Alturki and Nazik Alturki (Alturki & Alturki, 2021) conducted a study to forecast final grades and find honored students early on. They collected 300 undergraduate students’ records from one of the Saudi universities. They collected data from three departments: information technology, information science, and computer science. All of the participants are females and belong to the same age group (20–22). They used four alternative ways to examine the relevance of correlation attribute predictors. We discovered that 9 of the 18 proposed indicators for predicting students' academic progress after their fourth semester had a significant link. Student GPA over the first four semesters, the number of failed courses during the first four semesters, and the grades of three core courses, namely database fundamentals, computer network fundamentals, and programming language (1), are among these characteristics. Empirical results show that the main features that can predict students’ academic achievement are the grades of three core courses, the number of failed courses during the first four semesters, and the student's GPA during the first four semesters. Also, studying in a preparation year does not contribute to students’ success, and English language skills do not play an essential role in students’ success at the college of Computer and Information Sciences. Six classification algorithms were used: Nave Bayes, C4.5, Bayes Net with ADTree, Simple CART, Random Forest, and LADTree. After implementing those classifiers, the researcher found that Naive Bayes and Random Forest achieved the best results compared to other classifiers. The accuracy of Random Forest and Naive Bayes, respectively, was 92.6% and 85.8% after the third and fourth semesters.

Mustafa Yağcı (Yağcı, 2022) conducted a study to predict the academic performance of undergraduate students depending on midterm exam grades. The dataset for this study was taken from the student information system of a Turkish state university, which collects all student records. The dataset for these records includes the midterm and final exam grades, faculty, and department of 1854 students who studied the Turkish Language-I course in the 2019–2020 fall semester. The results of machine learning techniques such as support vector machines, random forests, logistic regression, nearest neighbor, k-nearest neighbor, and Naive Bayes were computed and compared to predict the students' final exam grades. Only three categories of parameters were used to make the predictions: midterm exam marks, department data, and faculty data. The results that were achieved in this study show the accuracy of the classifier at 70–75%.

Table 1 shows comparative among the previous studies depending on study the dataset, purpose of study, data mining techniques that used in study, and the results. As shown in Table 1, most of the studies achieved promising results, but the best results were obtained in the studies (Prabha & Shanavas, 2015) and (Christina, 2018) respectively. The best result gained from using MLP, J48 decision tree, Random forest, and SMO algorithms.

Table 1: Summary of Education field literature studies

| Reference | Dataset | Purpose of study | Techniques | Results | |
|---|---|---|---|---|---|
| (Prabha & Shanavas, 2015) | Taken from students' data who used the MATHSTUTOR e-learning tool. It contains 120 sixth-grade students | Assess student learning and identify weaknesses to improve learning outcomes | MLP | Accuracy | 100% |
| | | | J48 | | 100% |
| | | | Naive Bayes | | 98.33% |
| | | | Random Forest | | 96.66% |
| | | | ZeroR | | 29% |
| (Asif et al., 2017) | Collected from two academic batches using a sample of 210 undergraduate students | 1- Predicting pupils' academic progress at the end of a four-year program.<br>2- Examine typical progressions and putting them together with prediction results. Low-achieving and high-achieving pupils have been identified as two key groups. | Decision Tree with Information Gain | 60.58% | |
| | | | Decision Tree with Gini Index | 69.23% | |
| | | | Decision Tree with Accuracy | 55.77% | |
| | | | 1-Nearest Neighbour | 83.65% | |
| | | | Rule Induction with Information Gain | 74.04% | |
| | | | Naive Bayes | 62.50% | |
| | | | Random Forest with Gini Index | 69.23% | |
| | | | Neural Networks | 71.15% | |
| | | | Random Forest with Accuracy | 68.27% | |
| | | | Random Forest with Information Gain | 62.50% | |
| (Christina, 2018) | The dataset was gathered for approximately 350 understudies from the 4-stage B.E. CSE | predict test informative collections for future results, such as best, good, average, or poor | Naive Bayes | Accuracy | 85.92% |
| | | | MLP | | 94.94% |
| | | | SMO | | 94.34% |
| | | | Decision table | | 96.10% |
| | | | J48 | | 97.27% |
| | | | REP Tree | | 95.33% |
| (Manjula V & srinath, 2020) | 31 attributes for 5000 person | identify the reasons for students' weak performance through the use of psychology and sociology | AANN | Accuracy | 95% |
| | | | MLP | | ≈ 75% |
| | | | J48 | | ≈ 70% |
| | | | SMO | | ≈ 68% |
| | | | REP Tree | | ≈ 67% |
| | | | Naive Bayes | | ≈ 65% |
| (Alturki & Alturki, 2021) | 300 undergraduate students' records from one of the Saudi universities | Early on, forecast pupils' ultimate grades and recognize honorary students. | J48 | Accuracy | 77.6% |
| | | | LADTree | | 75.5% |
| | | | SimpleCart | | 79.6% |
| | | | Naive Bayes | | 77.6% |
| | | | Random Forest | | 92.6% |
| | | | Bayes Net with ADTree | | 79.6% |

| | | | | | |
|---|---|---|---|---|---|
| (Yağcı, 2022) | Includes the grades, faculty, and department of 1854 students who took the Turkish Language-I course in the autumn semester of 2019–2020 | predict the academic performance of undergraduate students depending on midterm exam grades | Random Forest | Accuracy | 74.6% |
| | | | Neural Network | | 74.6% |
| | | | SVM | | 73.5% |
| | | | Logistic Regression | | 71.7% |
| | | | Naive Bayes | | 71.3% |
| | | | kNN | | 69.9% |

### 3.1. In Fraud detection

Maria R. Lepoivre and et al. (Lepoivre et al., 2016) employed SIMPLEKMEANS and Principal Component Analysis unsupervised techniques to develop a credit card frauds system. The data contained five bank accounts. The first account has two legal transactions and one fraudulent transaction, while the second account has two fraudulent transactions and six legal transactions. There are 20 transactions in the third transaction, with 15% of them being fraudulent. The fourth transaction has three legal transactions, while the fifth account has 15 transactions with a fraud rate of 33.33 percent. To improve accuracy, the geographic positions of the client and transaction were added to the traditional collected data. The accuracy obtained from applying the proposed model to correctly classifying the transactions of bank accounts No. 1, 2, 4, and 5 is 100%.

Through the study to detect phishing emails and contrast the manual and automatic feature selection groups for email, Sa'id Abdullah (Al-Saaidah, 2017) built a new system using supervised and unsupervised techniques. In this study, Decision Tree (DT), Naive Bayes, Classification and Regression Trees, Logistic Regression, and Sequential Minimal Optimization (SMO) are some of the classification techniques that are reviewed and contrasted. The data that was used in this study was collected from two websites. The first was the monkey website for phishing emails, and the other one was the spam Assassin website for the data mining competition for legal emails. The dataset contains 4800 emails, including 2400 legitimate emails and 2400 phishing emails, representing the 47 features of the email structure. The study indicated that the best manually selected groups had an accuracy level of 98.25 percent, which was the same as the automatic features group. In addition, in both manual and automatic settings, the J48, Decision Tree, and SMO classifiers outperformed the other algorithms by providing a higher accuracy average. Furthermore, using the three top algorithms of SMO, Decision Tree, and J48, an integrated system of multiple classifiers was constructed, and the results showed that combining unsupervised and supervised techniques before testing improved the accuracy of detecting phishing emails, with a score of 98.37 for all features.

Navanshu Khare and Saad Yunus Sait (Navanshu Khare & Sait, 2018) used Random Forest, Decision Tree, Logistic Regression, and SVM techniques to detect fraud on credit cards. The dataset contained 284786 transactions that were collected from European cardholders. The results show that, for logistic regression, decision tree, Random Forest, and SVM classifiers, the best accuracy is 97.7%, 95.5 %, 98.6%, and 97.5 percent, respectively.

Kaithekuzhical Leena Kurien and Ajeet Chikkamannur (Kaithekuzhical Leena Kurien & Chikkamannur, 2019) applied Logistic Regression and a Random Forest classifier to detect fraudulent transactions in credit card usage. The dataset that was used in this study contained 284807 transactions that occurred over two days, including 492 frauds. Positive class (frauds) account for only 0.172 percent of all transactions, this dataset was heavily skewed. The accuracy for applying a Logistic Regression classifier was 88%, while the accuracy of Random Forest was 93%.

To predict the outcome of regular and fraudulent transactions, Maad M. Mijwil and Israa Ezzat Salem (Maad M. Mijwil & Salem, 2020) applied machine-learning classifiers (C4.5 decision trees, Naive Bayes, and Bagging Ensemble Learner). The dataset that was used in this work contained 297467 transactions via credit cards, including 3293 frauds. The performance result of the used algorithms was as follows: The C4.5 Decision Tree with 94.1% precision and 78.9% recall, the Bagging Ensemble with 91.6% precision and 80.7% recall, and the Naive Bayes classifier gave 65.6% precision and a recall of 81%.

Kayode Ayorinde (Ayorinde, 2021) employed Random Forest, Xgboost, Decision Trees, Logistic Regression, K-Means, and Neural Networks for credit card fraud prediction. The researcher used data for 1,000 credit card customers to do transactions. The total number of transactions in this dataset is 1048575, with 6006 fraudulent transactions recorded out of that total number of transactions. The dataset is severely skewed, with the positive class (frauds) accounting for only approximately 0.5727 percent of all transactions. The dataset comprises 22 features, including "Amount," "Category," "Is a Fraud," and others, that span multiple data types. The used algorithms such as Random Forest, Decision Trees, and Xgboost came out to be the best models that predict credit card fraud with AUC scores of 1.00%, 0.99%, and 0.99%, and the best accuracy for Random Forest, Xgboost Logistic, and Autoencoder with scores of 1.00%, 0.98%, and 0.98%, respectively.

Noor Saleh Alfaiz and Suliman Mohamed Fati (Noor Saleh Alfaiz & Fati, 2022) proposed a method for credit card fraud detection. The proposed scheme is split into two phases. The first phase aims to identify the top three machine learning algorithms from a pool of nine. The best three algorithms will be combined with nineteen resampling approaches in the second phase. Each approach in both phases was evaluated using AUC, F1-Score, accuracy, precision, and recall. The nine methods used in the first phase are: K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Gradient Boosting Machines (GBM), Category Boosting (CatBoost), and Extreme Gradient Boosting (XGBoost). The 19 resampling approaches are separated into 11 undersampling techniques, 6 oversampling techniques, and 2 combinations of both undersampling and oversampling techniques in the second phase. There are 66 models in total in both phases, with 330 evaluation metric values that took over a month to collect. The dataset that was used in this study contained 284807 transactions that occurred over two days, including 492 frauds. AllKNN, along with CatBoost, is the best model of all approaches (AllKNN-CatBoost). AllKNN-CatBoost was compared to earlier studies that used the same dataset and used comparable methods. Indeed, in terms of AUC (97.94%), recall (95.91%), and F1-Score, AllKNN-CatBoost beats earlier models (87.40%).

Table 2: Summary of fraud detection field literature studies

| Reference | Dataset | Purpose of study | Techniques | Results | |
|---|---|---|---|---|---|
| (Lepoivre et al., 2016) | 39 transaction from five bank accounts | develop a credit card frauds system | PCA and SIMPLEKMEANS | Accuracy | 100% |
| (Al-Saaidah, 2017) | 4800 emails, including 2400 legitimate emails and 2400 phishing emails from the monkey website for phishing emails and the spam Assassin website for legal emails | detect phishing emails and contrast the manual and automatic feature selection groups for email | Logistic Regression | Accuracy | 97.75% |
| | | | Decision Tree (DT, J48) | | 98% |
| | | | CART, One R | | 97% |
| | | | SMO | | 98% |
| | | | Naive Bayes | | 97.37% |

| | | | | | |
|---|---|---|---|---|---|
| (Navanshu Khare & Sait, 2018) | 284786 transactions that were collected from European cardholders | detect fraud on credit cards | Logistic Regression | Accuracy | 97.7% |
| | | | Decision Tree | | 95.5% |
| | | | Random Forest | | 98.6% |
| | | | SVM | | 97.5% |
| (Kaithekuzhical Leena Kurien & Chikkamannur, 2019) | 284807 transactions that occurred over two days, including 492 frauds | detect fraudulent transactions in credit card usage | Logistic Regression | Accuracy | 88% |
| | | | Random Forest | | 93% |
| (Maad M. Mijwil & Salem, 2020) | 297467 transactions including 3293 frauds | forecast the result of both legitimate and illegitimate transactions | | precision | recall |
| | | | C4.5 | 94.1% | 78.9% |
| | | | Bagging | 91.6% | 80.7% |
| | | | Naive Bayes | 65.6% | 81% |
| (Ayorinde, 2021) | 1048575, with 6006 fraudulent transactions | Credit card fraud prediction. | Random Forest | Accuracy | 100% |
| | | | XGBoost | | 98% |
| | | | Autoencoder | | 98% |
| (Noor Saleh Alfaiz & Fati, 2022) | 284807 transactions that occurred over two days, including 492 frauds | credit card fraud detection | Logistic Regression | Accuracy | 99.89% |
| | | | KNN | | 99.84% |
| | | | Decision Tree | | 99.99% |
| | | | Naive Bayes | | 99.29% |
| | | | Random Forest | | 99.96% |
| | | | GBM | | 99.90% |
| | | | LightGBM | | 99.59% |
| | | | XGBoost | | 99.96% |
| | | | Catboost | | 99.96% |

As shown in Table 2, most of the studies achieved good results, but the best results were obtained in the studies (Lepoivre et al., 2016), (Noor Saleh Alfaiz & Fati, 2022), (Ayorinde, 2021), (Al-Saaidah, 2017), and (Navanshu Khare & Sait, 2018) respectively. Also, the best algorithms that gave the best result of many studied were Random Forest, Decision Tree (DT, J48), Catboost, Xgboost, GBM, SMO, Logistic Regression, SVM, CART, and One R.

### 3.2. In Health care Field

A.S.Aneeshkumar and C.Jothi Venkateswaran (A.S.Aneeshkumar & Venkateswaran, 2015) used Snake peel-enabled hybrid Ant colony optimization and genetic algorithm, as well as fuzzy K-means for liver diseases classification. The dataset that was used in this study was collected from a reputed hospital in the southern region of Tamil Nadu. It contained 48 attributes and 6078 patients. The accuracy that was gained in this study for each kind of liver diseases was above 94%.

From the Liver Function Test (LFT) dataset, Tapas Ranjan Baitharua and Subhendu Kumar Panib (Tapas Ranjan Baitharu & Pani, 2016) conducted a study to predict liver diseases such as bile duct, cirrhosis, liver cancer, chronic hepatitis, and acute hepatitis. They used Naive Bayes, decision trees J48, ZeroR, multilayer perceptron, VFI, and the 1BK algorithm to classify liver diseases. After implementing the mentioned algorithms, the result showed that the performance of the Multilayer Perceptron was better than the others, with an accuracy of 71.59%.

Anamika Arora and Pradeep Chouksey (Anamika Arora & Chouksey, 2017) conducted a study to predict the presence of infertility in women. For predicting infertility, around sixteen attributes are required for detecting infertility or not. In this research work, sixteen attributes are reduced to eight attributes. Three classifiers like the Bagging algorithm, the J48 decision tree, and Naive Bayes are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of a number of attributes. The dataset was collected from different cities such as Aurangabad, Indore, Tatanagar, etc., and it contained 194 instances and 16 features. Then it was reduced to 11 features. The experimental results showed that the performance of the Bagging algorithm was better than the others, with an accuracy of 85.03%.

Mohammad Shafenoor Amin and et al. (Mohammad Shafenoor Amin et al., 2018) proposed a method to find relevant features and data mining techniques that can increase the accuracy of cardiovascular disease prediction. Naive Bayes, k-NN, Logistic Regression (LR), Decision Tree, Neural Network, Support Vector Machine (SVM), and Vote (a hybrid technique with Naive Bayes and Logistic Regression) were used to create prediction models using various combinations of features and the seven classification techniques. They used the Cleveland dataset, which contained 303 records and 76 attributes, the researchers implemented their proposal with 9 features only. The heart disease prediction model constructed using the identified significant features and the best-performing data mining technique (i.e., Vote) achieves an accuracy of 87.4% in heart disease prediction, according to the results of the experiments.

In the context of psychiatric patient datasets, E. Chandra Blessie and Bindu George (E. Chandra Blessie & George, 2019) compared the classifiers Naive Bayes and K-Nearest Neighbor. The data patterns that were retrieved may be valuable in preventing mental disease and assisting in the delivery of effective mental health treatments. They used a dataset including 65 instances and 36 features in it. Because the model's prediction accuracy was 90.71 percent, the Naive Bayes classification was able to construct the most accurate model. The Naive Bayes classification had a sensitivity and specificity of 76.50 percent and 86.71 percent, respectively. Model accuracies of 73.85 percent are produced by KNN classifiers.

In (Almustafa, 2020) Khaled Mohamad Almustafa used Naive Bayes, K-Nearest Neighbor (K-NN), JRip, Decision Tree J48, SVM, Adaboost, Decision Table (DT), and Stochastic Gradient Decent (SGD) algorithms for the prediction of heart diseases with minimal features. There were 76 features in the dataset for 1025 patients from Hungary, Cleveland, Long Beach, and Switzerland. Only 14 features from the set were used by the researcher. After these algorithms were implemented, the classification accuracy for the KNN (K=1), Decision Tree J48, and JRip classifiers was 99.7073, 98.0488, and 97.2683 percent, respectively. On the HD dataset, a feature extraction method was used, and the results showed improved performance in terms of classification accuracy for KNN (N=1) and Decision Table classifiers by 100 and 93.8537 percent, respectively, after using the selected features by only applying a combination of up to 4 attributes instead of 13 attributes for the HD case prediction.

Thyroid diseases were classified by Khalid Salman and Emrullah Sonuç (Khalid salman & Sonuç, 2021) into three types: hypothyroidism, hyperthyroidism, and normal. For classification, they used random forest, support vector machines, naive bayes, decision trees, k-nearest neighbors, logistic regression, linear discriminant analysis, and multilayer perceptron (MLP). The data was collected from external hospitals and laboratories specializing in analyzing and diagnosing diseases in Iraq. They collected 1250 records and 17 features for male and female patient's ages 1–90 years old. After reduction, 3 features of the Random Forest, Decision Tree, MLP (NN), SVM, Logistic Regression, KNN, Linear Discriminant Analysis, and Naive Bayes have an accuracy of 98.93%, 98.4%, 97.6%, 92.27%, 91.47%, 90.93%, 83.2%, and 81.33%, respectively.

Manav Mangukiya and et al. (Manav Mangukiya et al., 2022) diagnosed breast cancer in women. On the Breast Cancer Wisconsin Dataset from Kaggle, they employed Decision Tree, Support Vector Machine (SVM), K Nearest Neighbours (kNN), Naive Bayes (NB), Adaboost, Random Forest, and XGboost classification methods. It had 569 patient records and 32 features. According to the findings, XGboost has the highest accuracy (98.24 percent) and the lowest error rate.

Table 3: Summary of Health care field literature studies

| Reference | Dataset | Purpose of study | Techniques | Accuracy |
|---|---|---|---|---|
| (A.S.Aneeshkumar & Venkateswaran, 2015) | 48 attributes and 6078 patients from a reputed hospital in the southern region of Tamil Nadu | liver diseases classification | hybrid Ant colony optimization, genetic algorithm, and fuzzy K-means | 94% |
| (Tapas Ranjan Baitharu & Pani, 2016) | Liver Function Test (LFT) dataset | predict liver diseases such as bile duct, cirrhosis, liver cancer, chronic hepatitis, and acute hepatitis | J48 | 68.97% |
| | | | ZeroR | 57.97% |
| | | | MLP | 71.59% |
| | | | 1BK | 62.89% |
| | | | Naive Bayes | 55.36% |
| | | | VFI | 60.28% |
| (Anamika Arora & Chouksey, 2017) | 194 instances and 16 features collected from different cities | predict the presence of infertility in women | Naive Bayes | 82.31% |
| | | | J48 | 84.35% |
| | | | Bagging | 85.03% |
| (Mohammad Shafenoor Amin et al., 2018) | Cleveland dataset, which contained 303 records and 76 attributes | find relevant features and data mining techniques that can increase the accuracy of cardiovascular disease prediction | Vote | 87.41% |
| | | | Naive Bayes | 84.81% |
| | | | SVM | 85.19% |
| (E. Chandra Blessie & George, 2019) | 65 instances and 36 features | assisting in the delivery of effective mental health treatments | Naive Bayes | 90.71% |
| | | | KNN | 73.85% |
| (Almustafa, 2020) | 76 features in the dataset for 1025 patients from Hungary, Cleveland | prediction of heart diseases with minimal features | KNN (K=1) | 99.70% |
| | | | J48 | 98.04% |
| | | | JRip | 97.26% |
| (Khalid salman & Sonuç, 2021) | 1250 records and 17 features for male and female Iraqi patient's ages 1–90 years old | Thyroid diseases classification | Random Forest | 98.93% |
| | | | Decision Tree | 98.4% |
| | | | MLP | 97.6% |
| | | | SVM | 92.27% |
| | | | Logistic Regression | 91.47% |
| | | | KNN | 90.93% |
| | | | Linear Discriminant Analysis | 83.2% |
| | | | Naive Bayes | 81.33% |

| | | | | | |
|---|---|---|---|---|---|
| (Manav Mangukiya et al., 2022) | Wisconsin Dataset from Kaggle contained 569 patient records and 32 features | diagnosed breast cancer in women | | Without Standard scale | With Standard scale |
| | | | SVM | 57.89% | 96.49% |
| | | | KNN | 93.85% | 57.89% |
| | | | Random Forest | 97.36% | 75.43% |
| | | | Decision Tree | 94.73% | 75.43% |
| | | | Naive Bayes | 94.73% | 93.85% |
| | | | Adaboost | 94.73% | 94.73% |
| | | | XGboost | 98.24% | 98.24% |

After conducts a comparison in Table 3, the results shown that (Almustafa, 2020), (Khalid salman & Sonuç, 2021), (Manav Mangukiya et al., 2022), and (A.S.Aneeshkumar & Venkateswaran, 2015) get the best results, and the best algorithms that gave the best results in this field were KNN (K=1), Random Forest, XGboost, Decision Tree, MLP, the hybrid of Ant colony optimization, genetic algorithm, and fuzzy K-means, and Adaboost.

## 4. Conclusion

In this study, we highlighted some studies in the fields of education, fraud detection, and health care for the period from 2015 to February 2022. The techniques of data mining and machine learning have given promising results and contributed to facilitating work in many areas. During this research, it was observed that the technique's performance varied depending on the data set used. It is not possible to compare two works that used the same techniques but with a different data set, and vice versa. There are data mining techniques that have achieved promising results in some areas, but in some studies, there was a need to use more than one technique or hybrid techniques to obtain better results.

# References

[1] A.S.Aneeshkumar, & C. J. Venkateswaran, 2015, A novel approach for Liver disorder Classification using Data Mining Techniques, ***Engineering and Scientific International Journal (ESIJ)***, 2(1): 15-18.

[2] S. i. A. Al-Saaidah, 2017, ***Detecting Phishing Emails Using Machine Learning Techniques***, Middle East University.

[3] K. M. Almustafa, 2020, Prediction of heart disease and classifiers' sensitivity analysis, ***BMC Bioinformatics***, 21(1): 278.

[4] S. Alturki, & N. Alturki, 2021, Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions, ***Journal of Information Technology Education: Innovations in Practice***, 20: 121-137.

[5] Anamika Arora, & P. Chouksey, 2017, A Novel Approach for Women's Infertility Detection Using Data Mining Techniques, ***International Journal of Electronics Communication and Computer Engineering***, 8(2): 129-133.

[6] Anshu, 2019, Review Paper on Data Mining TechniquesandApplications, ***International Journal of Innovative Research in Computer Science & Technology***, 7(2): 22-26.

[7] Aruna, & P. K. Butey, 2014, Importance of Data Mining with Different Types of Data Applications and Challenging Areas, ***Int. Journal of Engineering Research and Applications***, 4(5): 38-41.

[8] R. Asif, A. Merceron, S. A. Ali, & N. G. Haider, 2017, Analyzing undergraduate students' performance using educational data mining, ***Computers & Education***, 113: 177-194.

[9] K. Ayorinde, 2021, ***A Methodology for Detecting Credit Card Fraud***, Minnesota State University, Mankato.

[10] M. Christina, 2018, Predicting Student Performance using Data Mining, ***International Journal of Computer Sciences and Engineering***, 6(10): 172-177.

[11] A. K. Dogra, & TanujWala, 2015, A Review Paper on Data Mining Techniques and Algorithms, ***International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)***, 4(5): 1976-1979.

[12] E. Chandra Blessie, & B. George, 2019, A Novel approach for Psychiatric Patient Detection and Prediction using Data Mining Techniques, ***International Journal of Engineering Research & Technology (IJERT)***, 7(5): 1-4.

[13] J. Han, & M. Kamber. 2006. Data Mining: Concepts and Techniques, second ed. United States of America: Diane Cerra.

[14] Julio Ponce, A. O. Alberto Hernández, A. P. Felipe Padilla, Francisco Álvarez,, & E. P. d. León. 2009. Data Mining in Web Applications. In J. P. a. A. Karahoca (Ed.), ***Data Mining and Knowledge Discovery in Real Life Applications***: 438. Austria: I-Tech.

[15] Kaithekuzhical Leena Kurien, & A. Chikkamannur, 2019, DETECTION AND PREDICTION OF CREDIT CARD FRAUD TRANSACTIONS USING MACHINE LEARNING, ***International Journal of Engineering Sciences &Research Technology***, 8(3): 199-208.

[16] Khalid salman, & E. Sonuç, 2021, Thyroid Disease Classification Using Machine Learning Algorithms, ***Journal of Physics: Conference Series***, 1963(1).

[17] M. R. Lepoivre, C. O. Avanzini, G. Bignon, L. Legendre, & A. K. Piwele, 2016, Credit Card Fraud Detection with Unsupervised Algorithms, ***Journal of Advances in Information Technology***, 7(1): 34-38.

[18] Maad M. Mijwil, & I. E. Salem, 2020, CreditCard Fraud Detectionin PaymentUsingMachine Learning Classifiers, ***Asian Journal of Computer and Information Systems***, 8(4): 50-55.

[19] Manav Mangukiya, Anuj Vaghani, & M. Savani, 2022, Breast Cancer Detection with Machine Learning, ***International Journal for Research in Applied Science and Engineering Technology***, 10(2): 141-145.

[20] Manjula V, & M. srinath, 2020, AN EFFECTIVE HIERARCICAL CLASSIFICATION IN EDUCATION DATA MINING TECHNIQUES TO VALUATE ENGINEERING STUDENTS PERFORMANCE (ADAPTIVE ARTIFICIAL NEURAL NETWORK), ***Journal of Seybold Report***, 15(8): 3009-3022.

[21] Mohammad Shafenoor Amin, Y. K. Chiam, & K. D. Varathan, 2018, Identification of significant features and data mining techniques in predicting heart disease, ***Telematics and Informatics***, 36: 82-93.

[22] Mustafa Abdalrassual Jassim, & S. N. Abdulwahid, 2021, Data Mining preparation: Process, Techniques and Major Issues in Data Analysis, ***IOP Conference Series: Materials Science and Engineering***, 1090(1).

[23] Muyeba, Maybin, M. S. Khan, & F. Coenen. 2009. Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework, ***New Frontiers in Applied Data Mining***: 49-61.

[24] Navanshu Khare, & S. Y. Sait, 2018, Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models, ***International Journal of Pure and Applied Mathematics***, 118(20): 825-837.

[25] Noor Saleh Alfaiz, & S. M. Fati, 2022, Enhanced Credit Card Fraud Detection Model Using Machine Learning, ***Electronics***, 11(4).

[26] P.Thangaraju, & D.Nanthini, 2015, AN EXHAUSTIVE STUDY ON ASSOCIATION RULE MINING, ***International Journal of Computer Science and Mobile Computing***, 4(3): 411 – 417.

[27] Pallavi Reddy, Ch. Mandakini, & C. Radhika, 2020, A Review on Data Mining Techniques and Challenges in Medical Field, ***International Journal of Engineering Research & Technology (IJERT)***, 9(8): 329-333.

[28] Prabha, S. Lakshmi, Shanavas, & A. R. Mohamed. 2014. Implementation of E-Learning Package for Mensuration-A Branch of Mathematics, ***2014 World Congress on Computing and Communication Technologies***: 219-221.

[29] S. L. Prabha, & A. R. M. Shanavas, 2015, Performance of Classification Algorithms on Students' Data – A Comparative Study, ***International Journal of Computer Science and Mobile Applications***, 3(9): 1-8.

[30] B. M. Ramageri, DATA MINING TECHNIQUES AND APPLICATIONS, ***Indian Journal of Computer Science and Engineering***, 1(4): 301-305.

[31] Sidhant Sethi, Dheeraj Malhotra, & N. Verma, 2016, Data Mining: Current Applications & Trends, ***International Journal of Innovations in Engineering and Technology (IJIET)***, 6(4): 667-673.

[32] Smita, & P. Sharma, 2014, Use of Data Mining in Various Field: A Survey Paper, ***IOSR Journal of Computer Engineering (IOSR-JCE)***, 16(3): 18-21.

[33] S. K. Solanki, & J. T. Patel. 2015. A Survey on Association Rule Mining, ***2015 Fifth International Conference on Advanced Computing & Communication Technologies***: 212-216.

[34] R. Tamilselvi, & S. Kalaiselvi, 2013, An Overview of Data Mining Techniques and Applications, ***International Journal of Science and Research***, 2(2): 506-509.

[35] Tapas Ranjan Baitharu, & S. K. Pani, 2016, Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset, ***Procedia Computer Science***, 85: 862-870.

[36] M. Yağcı, 2022, Educational data mining: prediction of students' academic performance using machine learning algorithms, ***Smart Learning Environments***, 9(1).