# A Survey On Various Data Dissemination Techniques Used In Data Analytics – Business Intelligence

M V Kamal[1], P Dileep[2]

[1] Research Scholoar, Dept of CSE, JNTU-Hyderabad, TS, India
[2] Research Scholoar, Dept of CSE, AU-Vishakapatanam, AP, India

**Abstract**— Business Intelligence (BI) are the set of strategies, processes, applications, data, technologies and technical architectures which are used to support the collection, data analysis, presentation and dissemination of business information. BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics and are capable of handling large amounts of structured and sometimes unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal is to allow for the easy interpretation of these big data.

**Index Terms**— Data Mining, Business Intelligence, Process Mining.

## 1. Introduction to Business Intelligence

Business Intelligence is a framework activity related to data mining that includes different applications, infrastructure and tools, and best practices that enable access to data analysis and information to improve and optimize decisions and performance [1].

Business intelligence combines a broad set of data analysis applications, including ad hoc analysis and querying, enterprise reporting, online analytical processing (OLAP)[2], mobile BI, real-time BI, operational BI, cloud and software as a service BI, open source BI, collaborative BI and location intelligence. BI applications can be bought separately from different vendors or as part of a unified BI platform from a single vendor.

Sporadic usage of the term business intelligence dates back as an umbrella category for applying data analysis techniques to support business decision-making processes [5][6]. What came to be known as BI technologies evolved from earlier, often mainframe-based analytical systems, such as decision support systems and executive information systems? Business intelligence is sometimes used interchangeably with business analytics; in other cases, business analytics is used either more narrowly to refer to advanced data analytics or more broadly to include both BI and advanced analytics [7].

Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability. Predicting student's performance becomes more challenging due to the large volume of data in educational databases. Currently in Malaysia, the lack of existing system to analyze and monitor the student progress and performance is not being addressed.

There are two main reasons of why this is happening. First, the study on existing prediction methods is still insufficient to identify the most suitable methods for predicting the performance of students in Malaysian institutions. Second is due to the lack of investigations on the factors affecting student's achievements in particular courses within Malaysian context. Therefore, a systematical literature review on predicting student performance by using data mining techniques is proposed to improve student's achievements.

The main objective of this paper is to provide an overview on the data mining techniques that have been used to predict student's performance. This paper also focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data. We could actually improve student's achievement and success more effectively in an efficient way using educational data mining techniques. It could bring the benefits and impacts to students, educators and academic institutions.

*BI solution technology consists of three primary components:*

- Data integration technology (Extract, transform, load) represents software that is used to analyze sources data, cleanup any data quality problem and transfer data in real time.
- Data warehouse Repositories (strategic & operational) represents the computer's hardware where extracted data is stored.
- BI tools – represents software tools that are used to model and analyze data to support better decisions.

*Various application area of BI that can be used in mining the data like:*

1. Analyzing the students behavior
2. Tracking the various systems

3. Financial planning of budget
4. Optimization of the processes
5. Web analytics
6. E-Commerce applications
7. Risk analysis
8. Relationship with customers and managing them etc.

*BI basic elements include the following:*

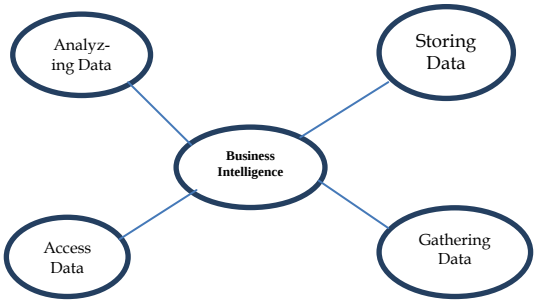



Figure1: Components of Business Intelligence.

*The functions and roles typically associated with the development of BI applications include the following:*

- ***Target decisions:*** Identify the key business decisions for which BI is needed.
- ***Target data sources:*** Identify data sources that can provide the source data needed.
- ***Extract, transform and load source data:*** Analyze the source data to determine requirements for extracting, transforming and transferring data for subsequent analytics in real time.
- ***Decision support modeling:*** Develop the logic, models and display formats by which the source data can be analyzed, mined, correlated, mapped, displayed and reported.
- ***Analysis and reporting:*** Develop the actual reports, queries, graphs, dashboards, or scoreboards to be used analyze and display or report the data in the data warehouse.

## 2. Next Generation Business Intelligence Application Development Technologies (NGBIADT)

Enterprises today are dealing with Next-generation Business Intelligence offerings, enable insights driven methods [4]. The cutting-edge technologies for advanced analytics are also becoming a challenge future in the areas of data mining.

*These include the following the themes:*

1. Customer Insights
2. Data monetization
3. Operational efficiency
4. Risk management and planning

In order to solve these themes related to data analytics, the developer should undergo the tedious process involved in the areas of data mining, which can be applied through code blocks[8][10]. lots of technologies were invented to improve the evaluation performance of any category of data to be analyzed. Business Intelligence is to get proper knowledge to make intelligent decisions but most BI systems are partial or local, which lack systematic information compilation and processing therefore compilation of extra information can lead to information and knowledge excess which make business intelligence difficult. Usually the business intelligence software scale is massive; Developing Business Intelligence Framework to Automate Data Mapping, Validation, and Data Loading from User Application, maintenance cycle is long, and costly. It is difficult for the small and medium-size companies to develop BI system.

The problem is with data model blueprints and ontology style that face amalgamation of business intelligence systems based on different data sources and structure. Integration framework for business intelligence system (BIS) and services resources based-ontology is proposed as well as data model prototypes and ontology method which can resolve the different data sources and structures integration dilemma. BIS assimilation based on ontology is the central part of the architecture is the repository; this stores configuration information about the IT infrastructure, the metadata for all applications, projects, scenarios, and execution logs. Repositories can be installed on an OLTP rational database. This metadata is stored in a centralized metadata repository.

There is only one master repository, which holds the following information:
1) Safety information together with users, profiles and access privileges for the data integration units.
2) Topology information together with technologies, definitions of servers and schemas, contexts and languages.
3) Old edition of objects. The information controlled in the master repository is maintained with Topology Manager and Security Manager.

## 3. Techniques Used

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information [9]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is also defined as technique used in extracting useful information from a large database. There are many data mining techniques like classification,

regression, clustering. Recently, data mining is used in educational area. It is known as Educational Data Mining (EDM). Mainly to predict student performance we use classification and regression techniques.

1. Classification
2. Regression

## 1. Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. A classification task begins with a data set in which the class assignments are known. Classifications are discrete and do not imply order[11]. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

As classification is used for prediction. There are many algorithms used to predict the things which we require. Algorithms used are Decision tree [12], Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine. Each algorithm has its own prerequisites and also they give accurate results.

### *Classification and Prediction Issues:*

*The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities:*

**Data Cleaning:** Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

**Relevance Analysis:** Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.

**Data Transformation and reduction:** The data can be transformed by any of the following methods.

**Normalization:** The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.

**Generalization:** The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

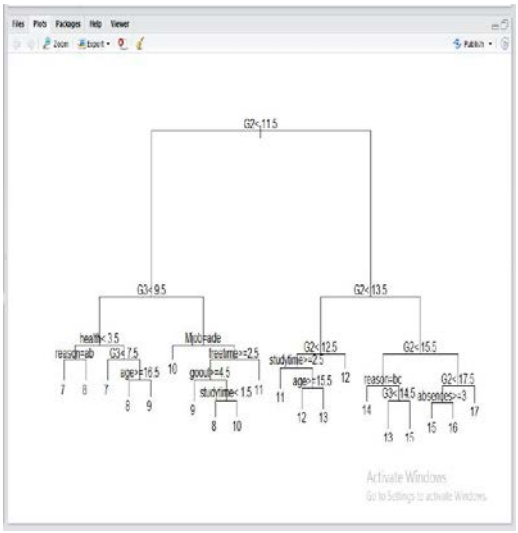
Data can also be reduced by some other methods such as wavelet transformation, binning, histogram analysis, and clustering.

### *Comparison of Classification and Prediction Methods:*

*Criteria for comparing the methods of Classification and Prediction is as listed below:*

1. **Accuracy:** Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
2. **Speed:** This refers to the computational cost in generating and using the classifier or predictor.
3. **Robustness:** It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
4. **Scalability:** Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.
5. **Interpretability:** It refers to what extent the classifier or predictor understands.

### Classification Sample Example:



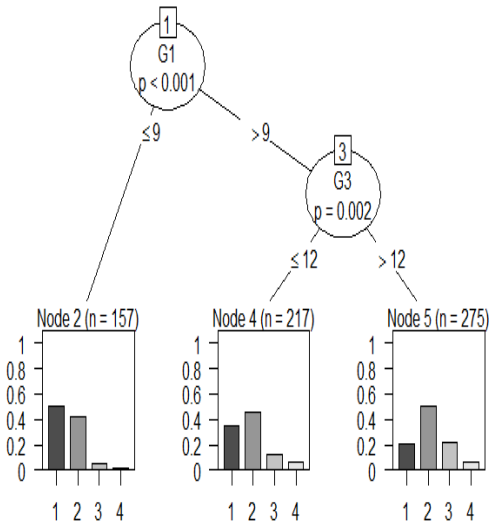

## 4. Algorithms Classified

1. Decision Tree
2. Neural Network
3. Naive Bayes
4. K-Nearest Neighbor
5. Support Vector Machine

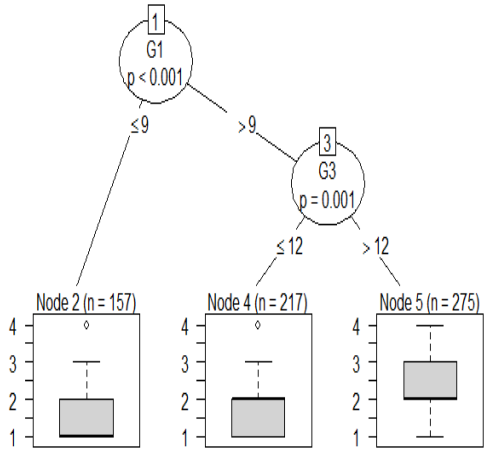
### Decision Tree

Decision Tree is one of a popular technique for prediction. Most of researchers have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the value. Rome roet al. said that the decision tree models are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules. There are approximately ten papers that have used Decision Tree as their method to evaluate students' performance. Examples of previous studies using Decision Tree method are predicting drop out features of student's data for academic performance, predicting third semester performance of MCA students and also predicting the suitable career for a student through their behavioral patterns. The students' performance evaluation is based on features extracted from logged data in an education web-based system. The examples of dataset are student's final grades, final cumulative grade point average (CGPA) and marks obtained in particular courses. All this datasets were studied and analyzed to find out the main attributes or factors that may affects the students' performance. Then, the suitable data mining algorithm will be investigated to predict students' performance. Mayilvaganan and Kapalna devi, have compared the classification techniques for predicting students' performance in their study. Meanwhile, Gray et al. Investigated the accuracy of classification models to predict learner's progression in tertiary education.

**Sample Decision Tree_1:**




**Sample Decision Tree_2:**




**Neural Network**

Neural network is another popular technique used in educational data mining. The advantage of neural network is that it has the ability to detect all possible interactions between predictors variables .Neural network could also do a complete detection without having any doubt even in complex nonlinear relationship between dependent and independent variables. Therefore, neural network technique is selected as one of the best prediction method. Through the meta-analysis study, eight papers have been published using Neural Network method. The papers present an Artificial Neural Network model to predict students' performance [13]. The attributes analyzed by Neural Network are admission data, student's attitude towards self-regulated learning and academic performance. The rest are same papers in addition with Decision Tree method where researchers have used both techniques to compare which one is the best prediction method for analyzing students' performance.

**Naive Bayes**

Naive Bayes algorithm is also an option for researchers to make a prediction. Among thirty papers, there are four papers that have used Naive Bayes algorithms to estimate student's performance. The objective of all these four papers is to find the most effective prediction technique [12] in predicting student's performance by making comparisons. Their research showed that Naive Bayes has used all of attributes contained in the data. Then, it analyzed each one of them to show the importance and independency of each attributes.

**K Nearest Neighbor**

As depicts in Table 5, all three papers studied in this research showed that K-Nearest Neighbor gave the best performance with the good accuracy. According to Bigdoli et al. (2003), K-Nearest Neighbor method had taken less time

to identify the students' performance as a slow learner, average learner, good learner and excellent learner. K-Nearest Neighbor gives a good accuracy in estimating the detailed pattern for learner's progression in tertiary education.

**Support Vector Machine**

Support Vector Machine is a supervised learning method used for classification. There are three papers that have used Support Vector Machine as their method to predict students' performance. Hamalainen et al. had chosen Support Vector Machine as their prediction technique because it suited well in small datasets [10]. Sembiring et al. stated that Support Vector Machine has a good generalization ability and faster than other methods. Meanwhile, the study done by Gray et al demonstrated that Support Vector Machine method [14] has acquired the highest prediction accuracy in identifying students at risk of failing. It shows the result of prediction accuracy.

**2. Regression**

Regression is widely used for prediction and forecasting in field of machine learning. Focus of regression is on the relationship between dependent and one or more independent variables [15][16].

Regression is a data mining or machine learning technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line(y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Basically a Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted, the factor that the equation solves for is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables.

## 5. Conclusion

The importance to Business Intelligence and the methodologies towards Next Generation Business Intelligence Application Development Technologies has been discussed with its various application areas. The use of data mining in enrollment management is a fairly new development. Current data mining is done primarily on simple numeric and categorical data. In the future, data mining will include more complex data types.Predicting students' performance is mostly useful to help the educators and learners improving their learning and teachingprocess. This paper has reviewed previous studies on predicting students' performance with various analytical methods. Most of the researchers have used cumulative grade point average (CGPA) and internal assessment as data sets. While for prediction techniques, the classification method is frequently used in educational data mining area. Under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting students'performance. In this paper, three supervised data mining algorithms were applied on the preoperative assessment data to predict success in a course (either passed or failed) and the performance of the learning methods were evaluated based on their predictive accuracy, ease of learning and user friendly characteristics. The results indicate that the Naïve Bayes classifier outperforms in prediction decision tree and neural network methods. It has also been indicated that a good classifier model has to be both accurate and comprehensible for professors. This study was based on traditional classroom environments, since the data mining techniques were applied after the data was collected.

## 6. Future Enhancement

The system can be enhanced with more distinctive attributes to get more accurate results, useful to improve the students learning outcomes. Also, experiments could be done using other data mining algorithms to get a broader approach, and more valuable and accurate outputs. Some different software may be utilized while at the same time various factors will be used.

## 7. References

[1]. Tao Xie, Suresh Thummalapenta, David lo, Chao Liu, "Data Mining for Software Engineering", IEEE Computer, Auguat 2009, pp. 55-62.

[2]. Hamid Abdul BAsit, Stan Jarzabek, " A Data Mining approach for detecting higher-level clones in Software", IEEE Transactions on Software Engineering, Vol. 35, No. 4, July/August 2009, pp. 497-514.

[3]. IvanoMalavelta, Henry Muccini, PatrizioPellicciona, Damien Andrew Tamburri, "Providing Architectural Languages and Tools Interoperability through Model Transformation Technologies", IEEE Transactions on Software Engineering, Vol. 36, No. 4, January/February 2010, pp. 119-140.

[4] Baker, Ryan Shaun, and Paul Salvador Inventado. "Educational data mining and learning analytics." In Learning analytics, pp. 61-75. Springer New York, 2014.

[5] Chamizo-Gonzalez, Julian, Elisa Isabel Cano-Montero, Elena Urquia-Grande, and Clara Isabel Muñoz-Colomina. "Educational data mining for improving learning outcomes in teaching accounting within higher education." The International Journal of Information and Learning Technology 32, no. 5 (2015): 272-285.

[6] Romero, C., Ventura, S.: Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3(1), 12–27, 2013

[7] Dawson, S., Gašević, D., Siemens, G., &Joksimovic, S.

(2014). Current State and Future Trends: A Citation Network Analysis of the Learning Analytics Field. In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge (pp. 231–240). New York, NY, USA: ACM.

[8] Yukselturk E, Ozekes S, Türel YK. Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and E-learning. 2014 Jul 1;17(1):118-33.

[9] Barber, R., & Sharkey, M. Course correction: Using analytics to predict course success. In Proceedings of the 2nd international conference on learning analytics and knowledge , 2012 ,pp. 259–262. ACM.

[10] Wolff, A., Zdrahal, Z., Nikolov, A., &Pantucek, M. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In Proceedings of the third international conference on learning analytics and knowledge, 2013, pp. 145–149. ACM

[11] A. Sharma, R. Kumar, P. K. Varadwaj, A. Ahmad, and G. M. Ashraf, "A comparative study of support vector machine, artificial neural network and bayesian classifier formutagenicity prediction," Interdisciplinary Sciences, Computational Life Sciences, vol. 3, no. 3, pp. 232–239, 2011.

[12] B. K. Bhardwaj and S. Pal, "Data mining: a prediction for performance improvement using classification," International Journal of Computer Science and Information Security, vol. 9, no.4, pp. 1–5, 2011.

[13] S.Huang and N.Fang, "Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models," Computers & Education, vol. 61, pp.133–145.

[14] R. Singh, A. Kainthola, and T. N. Singh, "Estimation of elastic constant of rocks using an ANFIS approach," Applied Soft Computing Journal, vol. 12, no. 1, pp. 40–45, 2012.

[15] J.-S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," IEEE Transactions on Systems, Man and Cybernetics, vol. 23, no. 3, pp. 665–685, 1993.

[16] K. S. Pal and S. Mitra, Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing, JohnWiley& Sons, New York, NY, USA, 1999.

[17] Jai Ruby & K. David, "Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms – A Case Study ", IJRASET International Journal for Research in Applied Science & Engineering Technology, Volume 2 Issue XI, November 2014

[18] Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R., &Alajrami, E. Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. International Journal of Hybrid Information Technology, 8(2),pp 221-228, 2015.

[19] Ajith P,,Tejaswi B, Sai MSS. Rule Mining Framework for Students Performance Evaluation. International Journal of Soft Computing and Engineering. 2013; 2(6):201–6.

[20] Kamber, Jiawei Han &Micheline. Data Mining: Concepts and Techniques second edition. San Francisco : Morgan Kaufmann, 2006.

[21] Jang, J. S. R., ANFIS: adaptive-network based fuzzy inference system. IEEE Trans. Syst., Man, Cybernetics, 1993, 23(3), 665– 685.

[22] M. Alizadeh, R. Rada, A.K.G. Balagh, M.M.S. Esfahani, Forecasting Exchange Rates: A Neuro-Fuzzy Approach, IFSA-EUSFLAT, 2009, pp.1745-1750.