INTRODUCTION TO DATA MINING AND PROCESS LEARNING TECHNIQUES

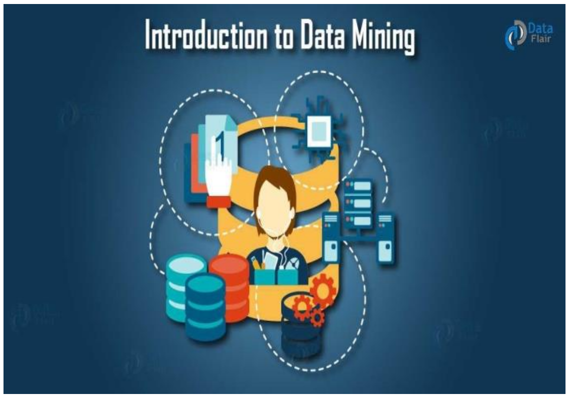
A.PALANIAMMAL
HEAD,DEPT OF BCA
SREE ARUMUGAM ARTS AND SCIENCE COLLEGE
THOLUDUR, TAMILNADU, INDIA

Abstract:

The data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

Keywords:Data mining, KDD, ANN, PROCESS LEARNING TECHNIQUES, Data mining process and architecture

1. INTRODUCTION



Data Mining is a set of method that applies to large and complex databases. This is to eliminate the randomness and discover the hidden pattern. As these data mining methods are almost always computationally intensive. We use data mining tools, methodologies, and theories for revealing patterns in data. There are too many driving forces present. And, this is the reason why data mining has become such an important area of study.

1.1 Data Mining History

In 1960s statisticians used the terms “Data Fishing” or “Data Dredging”. That was to refer what they considered the bad practice of analyzing data. The term “Data Mining” appeared around 1990 in the database community.

1.2 Foundation of Data Mining

We use data mining techniques for a long process of research and product development. As this evolution was started when business data was first stored on computers. Also, it allows users to navigate through their data in real time. We use data mining in the business community because it is supported by three technologies that are now mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

1.3 Type of data gathered

a. Business transactions

In this business industry, every transaction is “memorized” for perpetuity. We can say many transactions are dealing with time and can be inter-business deals such as purchases, exchanges, banking, stock, etc.,

b. Scientific data

Everywhere, our society is amassing colossal amounts of scientific data. As that scientific data need to be analyzed. Unfortunately, we have to capture and store more new data faster. Then we can analyze the old data already accumulated.

c. Medical and personal data

As we can say from the government to customer and for personal needs, we have to gather large information. That information is required for individuals and groups.

When correlated with other data, this information can shed light on customer behavior.

d. Surveillance video and pictures

As with the collapse of video camera prices, video cameras are becoming ubiquitous. Also, we can recycle cameras, videotapes from surveillance. However, it’s become a trend to store the tapes and even digitize them for future use and analysis.

e. Games

In societies, a huge amount of data and statistics is used. That is to collect about games, players, and athletes. As this information data is used by commentators and journalists for reporting.

f. Digital media

There are too many reasons for causes of the explosion in digital media repositories. Such as cheap scanners, desktop video cameras, and digital cameras. Associations such as the NHL and the NBA. That have already started converting their huge game collection into digital forms.

g. CAD and Software engineering data

There are multiple CAD systems for architects present to design building. As these systems are used to generate a huge amount of data.

Moreover, we can use S.E is a source of considerable similar data with code and objects that needs to be powerful tools for management and maintenance.

h. Virtual Worlds

Nowadays many applications are using three-dimensional virtual spaces. Also, these spaces and the objects they contain have to describe with special languages such as VRML. Ideally, we have to define virtual spaces as they can share objects and places. Also, there present the remarkable amount of virtual reality object available.

i. Text reports and memos (e-mail messages)

As communications are based on the reports and memos in textual forms in many companies. As they are exchanged by e-mail. Although, we use to store it in digital form for future use. Also, reference creating formidable digital libraries.

1.4 Uses of Data Mining

a. Automated prediction of trends and behaviors

We use to automate the process of finding predictive information in large databases. Questions that required extensive hands-on analysis can now be answered from the data. Targeted marketing is a typical example of predictive marketing. As we also use data mining on past promotional mailings. That is to identify the targets to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default. And identifying segments of a population likely to respond similarly to given events.

b. Automated discovery of previously unknown patterns

As we use data mining tools to sweep through databases. Also, to identify previously hidden patterns in one step. There is a very good example of pattern discovery. As it is the analysis of retail sales data. That is to identify unrelated products that are often purchased together. Also, there are other pattern discovery problems. That includes detecting fraudulent credit card transactions. It is identified that anomalous data could represent data entry keying errors.

1.4 Why Data Mining

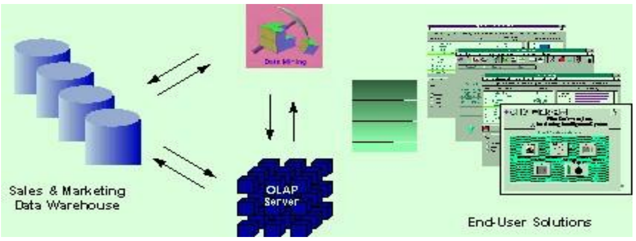
As data mining is having spacious applications. Thus, it is the young and promising field for the present generation. It has attracted a great deal of attention in the information industry and in society. Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, we use information and knowledge for applications ranging from market analysis. This is the reason why data mining is also called as knowledge discovery from data.

1.5 Data Mining Architecture

We need to apply advanced techniques in the best way. As they must be fully integrated with a data business analysis tools. To operate data mining tools we need extra steps for the extracting, and importing the data.

Furthermore, new insights need operational implementation, integration with the warehouse simplifies the application. We have to apply analytic data warehouse to improve business processes. Particularly in areas such as promotional campaign management, and so on.

Below figure illustrates an architecture for advanced analysis in a large data warehouse.



The ideal starting point is a data warehouse that must contain a combination of internal data tracking all customer contact. This should coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. Although, this warehouse can be implemented in a variety of relational database systems.

Such as Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model. That need to apply when navigating the data warehouse. Although, multidimensional structures allow the user to analyze the data. As they want to view their business. Such as summarizing by product line, region.

Further, the Data Mining Server must be integrated with the data warehouse. And, the OLAP server to embed ROI-focused business analysis directly into this infrastructure. Also, integration with the data warehouse enables the operational decisions. That is to be implemented and tracked.

Also, keep warehouse grows with new decisions and results. Thus, the organization can mine the best practices and apply them to future decisions

In the OLAP, results enhance the metadata. That is by providing a dynamic metadata layer. As this layer is used to represents a distilled view of the data. Reporting, visualization, and tools can then be applied to plan future actions. And confirm the impact of those plans.

1.6 Data Mining Process

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD). Also, nontrivial extraction of implicit information from data in databases.

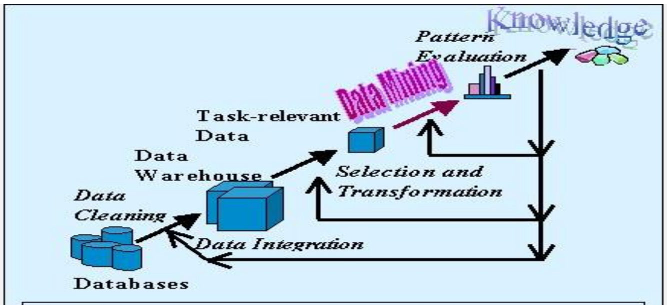


Figure 1.1: Data Mining is the core of Knowledge Discovery process

This process comprises of a few steps. That is to lead from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

a. Data cleaning

This is also called as data cleansing. As in this phase noise data and irrelevant data are removed from the collection.

b. Data integration

In this multiple data is combined at the same place.

c. Data selection

It decide the data relevant to the analysis is decided on and retrieved from the data collection.

d. Data transformation

It is also a data consolidation method. Also, it’s a phase in which the selected data is transformed into forms. That are appropriate for the mining procedure.

e. Data mining

In this, we have to apply clever techniques to extract patterns potentially useful.

f. Pattern evaluation

In this process interesting patterns representing knowledge are identified based on given measures.

g. Knowledge representation

It is the final phase. Particularly in this phase, knowledge is discovered and represented to the user. This essential step uses visualization techniques. That help users understand and interpret the data mining results.

1.7 Categories of Data Mining Systems

As there are too many data mining systems available. Also, some systems are specific that we need to dedicate to a given data source. Further, according to various criteria, data mining systems have to categorize.

a. Classification according to the type of data source mined

According to the type of data handle, have to perform classification of data mining. Such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

b. Classification according to the data model drawn on

In this classification is done on the basis of a data model. Such as relational database, object-oriented database, data warehouse, transactional, etc.

c. Classification according to the king of knowledge discovered

In this classification it is been done on the basis of the kind of knowledge. Such as characterization, discrimination, association, classification, clustering, etc.

d. Classification according to mining techniques used

As data mining systems employ are used to provide different techniques. According to the data analysis, we have to done this classification. Such as machine learning, neural networks, genetic algorithms, , etc.

1.8 Applications of Data Mining

- Weather forecasting.
- E-commerce.
- Self-driving cars.
- Hazards of new medicine.
- Space research.
- Fraud detection.
- Stock trade analysis.
- Business forecasting.
- Social networks.
- Customers likelihood.

More applications inlcude:

- A credit card company can leverage its vast warehouse of customer transaction data. As we perform this to identify customers. It shows more interest in a new credit product.

- Moreover, we use small test mailing. So the attributes of customers with an affinity for the product have to identify. Recent projects have indicated more than a 20-fold decrease in costs. That is target for mailing campaigns over conventional approaches.

- As diversified transportation company used to apply data mining. That is to identify the best prospects for its services. Further, need to apply this segmentation to a general business database. Such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

- Large consumer packaged goods company. That can apply data mining to improve its sales process to retailers. Although, data from consumer panel, and competitor activity have to apply. That is to understand the reasons for brand and store switching.

- Through this analysis, we have to manufacturer it. Then select promotional strategies that best reach their target customer segments.

1. RELATED WORKS:

2.1 Data Mining Techniques

a. Artificial neural networks

We use data mining in non-linear predictive models. As this learn through training and resemble biological neural networks in structure.

b. Decision trees

As we use tree-shaped structures to represent sets of decisions. Also, by this rules are generated for the classification of a dataset. These decisions generate rules for the classification of a dataset. As there are specific decision tree methods that includes Classification and Regression Trees and Chi-Square Automatic Interaction Detection (CHAID).

c. Genetic algorithms

There are the present genetic combination, mutation, and natural selection for optimization techniques. That is design based on the concepts of evolution.

d. Nearest neighbor method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) like. it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

e. Rule induction

The extraction of useful if-then rules from data based on statistical significance.

2.2 Data Mining Terminologies

a. Notation

Input X: X is often multidimensional.

Each dimension of X is denoted by Xj and is referred to as a feature variable or , variable.

Output Y: called the response or dependent variable.

A response is available only when learning is supervised.

b. Nature of Data Sets

a. Quantitative: Measurements or counts, recorded as numerical values, e.g. Height, Temperature, # of Red M&M's in a bag

b. Qualitative: Group or categories

c. Ordinal: possesses a natural ordering, e.g. Shirt sizes (S, M, L, XL)

d. Nominal: just name of the categories, e.g. Marital Status, Gender,

Color of M&M's in a bag

2. METHODOLOGIES:

Data mining is the process of looking at large banks of information to generate new information. Intuitively, you might think that data "mining" refers to the extraction of new data, but this isn't the case; instead, data mining is about extrapolating patterns and new knowledge from the data you've already collected.

Relying on techniques and technologies from the intersection of database management, statistics, and machine learning, specialists in data mining have dedicated their careers to better understanding how to process and draw conclusions from vast amounts of information. But what are the techniques they use to make this happen?

4.1 Data Mining Techniques

Data mining is highly effective, so long as it draws upon one or more of these techniques:

1. Tracking patterns. One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

2. Classification. Classification is a more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

3. Association. Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

4. Outlier detection. In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

5. Clustering. Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.

6. Regression. Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

7. Prediction. Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

academic record of students and course records.

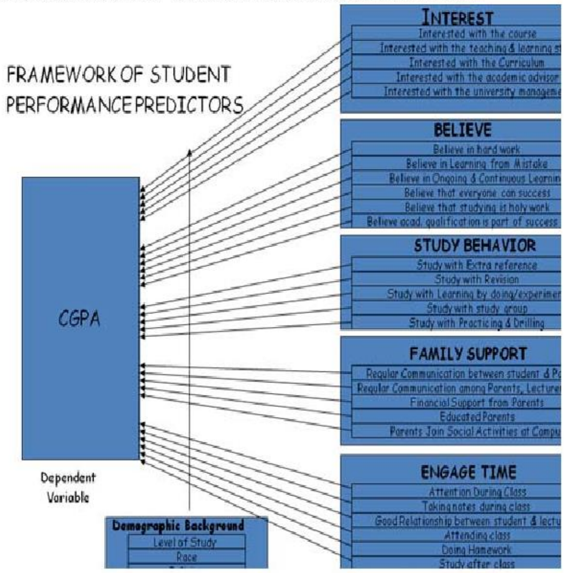

FRAMEWORK OF STUDENT PERFORMANCE PREDICTORS
CGPA
Dependent Variable
INTEREST
Interested with the course
Interested with the teaching & learning s
Interested with the Curriculum
Interested with the academic advisor
Interested with the university manageme
BELIEVE
Believe in hard work
Believe in Learning from Mistake
Believe in Ongoing & Continuous Learnin
Believe that everyone can success
Believe that studying is holy work
Believe acad. qualification is part of success
STUDY BEHAVIOR
Study with Extra reference
Study with Revision
Study with Learning by doing/experimen
Study with study group
Study with Practicing & Drilling
FAMILY SUPPORT
Regular Communication between student & Pa
Regular Communication among Parents, Lecture
Financial Support from Parents
Educated Parents
Parents Join Social Activities at Campu
ENGAGE TIME
Attention During Class
Taking notes during class
Good Relationship between student & lectu
Attending class
Doing Homework
Study after class
Demographic Background
Level of Study
Race

Figure 1. Framework of Student Performance Predictors

5. CONCLUSION*:*

The data mining is a knowledge – driven process and all stages contribute to the success of the process.if the domain experts play significant role in most phases of the process.the data mining need for selection of algorithms and techniques that support interpretection of mined knowledge, then expect for integrated tools and adequate techniques to support involvment of domain in the process.

3. PERFORMANCE ANALYSIS

REFERENCES:

1. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2011. [Google Scholar]

2. Kotsiantis, S.; Pierrakeas, C.; Pintelas, P. Predicting Students' Performance in Distance Learning using Machine Learning Techniques. *Appl. Artif. Intell.* **2004**, *18*, 411–426. [Google Scholar] [CrossRef]

3. Navamani, J.; Kannammal, A. Predicting performance of schools by applying data mining techniques on public examination results. *Res. J. Appl. Sci. Eng. Technol.* **2015**, *9*, 262–271. [Google Scholar] [CrossRef]

4. Moseley, L.; Mead, D. Predicting who will drop out of nursing courses: A machine learning exercise. *Nurse Educ. Today* **2008**, *28*, 469–475. [Google Scholar] [CrossRef] [PubMed]

5. Nandeshwar, A.; Menzies, T.; Nelson, A. Learning patterns of university student retention. *Expert Syst. Appl.* **2011**, *38*, 14984–14996. [Google Scholar] [CrossRef]

6. Thammasiri, D.; Delen, D.; Meesad, P.; Kasap, N. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Syst. Appl.* **2014**, *41*, 321–330. [Google Scholar] [CrossRef]

7. Dewan, M.; Lin, F.; Wen, D.; Kinshuk. Predicting dropout-prone students in e-learning education system. In Proceedings of the 2015 IEEE 12th Intl Conference on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conference on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), Beijing, China, 14 August 2016; pp. 1735–1740. [Google Scholar] [CrossRef]

8. Tan, M.; Shao, P. Prediction of student dropout in E-learning program through the use of machine learning method. *Int. J. Emerg. Technol. Learn.* **2015**, *10*, 11–17. [Google Scholar] [CrossRef]

9. Sultana, S.; Khan, S.; Abbas, M. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *Int. J. Electr. Eng. Educ.* **2017**, *54*, 105–118. [Google Scholar] [CrossRef]

10. Chen, Y.; Johri, A.; Rangwala, H. Running out of STEM: A comparative study across STEM majors of college students At-Risk of dropping out early. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge, Sydney, Australia, 7–9 March 2018; pp. 270–279. [Google Scholar]

11. Nagy, M.; Molontay, R. Predicting Dropout in Higher Education Based on Secondary School Performance. In Proceedings of the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria, Spain, 21 June 2018; pp. 000389–000394. [Google Scholar] [CrossRef]

12. Serra, A.; Perchinunno, P.; Bilancia, M. Predicting student dropouts in higher education using supervised classification algorithms. *Lect. Notes Comput. Sci.* **2018**, *10962 LNCS*, 18–33. [Google Scholar] [CrossRef]

13. Gray, C.; Perkins, D. Utilizing early engagement and machine learning to predict student outcomes. *Comput. Educ.* **2019**, *131*, 22–32. [Google Scholar] [CrossRef]

14. Chung, J.; Lee, S. Dropout early warning systems for high school students using machine learning. *Child. Youth Serv. Rev.* **2019**, *96*, 346–353. [Google Scholar] [CrossRef]

15. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.; Ali, S. Using machine learning to predict student difficulties from learning session data. *Artif. Intell. Rev.* **2018**, *52*, 1–27. [Google Scholar] [CrossRef]

16. Huang, S.; Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Comput. Educ.* **2013**, *61*, 133–145. [Google Scholar] [CrossRef]

17. Slim, A.; Heileman, G.L.; Kozlick, J.; Abdallah, C.T. Predicting student success based on prior performance. In Proceedings of the 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Singapore, 16 April 2015; pp. 410–415. [Google Scholar] [CrossRef]

18. Zhao, C.; Yang, J.; Liang, J.; Li, C. Discover learning behavior patterns to predict certification. In Proceedings of the 2016 11th International Conference on Computer Science & Education (ICCSE), Nagoya, Japan, 23 August 2016; pp. 69–73. [Google Scholar] [CrossRef]

19. Chaudhury, P.; Mishra, S.; Tripathy, H.; Kishore, B. Enhancing the capabilities of student result prediction system. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, Udaipur, India, 4–5 March 2016. [Google Scholar] [CrossRef]

20. Nespereira, C.; Elhariri, E.; El-Bendary, N.; Vilas, A.; Redondo, R. Machine learning based classification approach for predicting students performance in blended learning. *Adv. Intell. Syst. Comput.* **2016**, *407*, 47–56. [Google Scholar] [CrossRef]

21. Sagar, M.; Gupta, A.; Kaushal, R. Performance prediction and behavioral analysis of student programming ability. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21 September 2016; pp. 1039–1045. [Google Scholar] [CrossRef]

22. Verhun, V.; Batyuk, A.; Voityshyn, V. Learning Analysis as a Tool for Predicting Student Performance. In Proceedings of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 11 September 2018; Volume 2, pp. 76–79. [Google Scholar] [CrossRef]

23. Backenköhler, M.; Wolf, V. Student performance prediction and optimal course selection: An MDP approach. *Lect. Notes Comput. Sci.* **2018**, *10729 LNCS*, 40–47. [Google Scholar] [CrossRef]

24. Hsieh, Y.Z.; Su, M.C.; Jeng, Y.L. The jacobian matrix-based learning machine in student. *Lect. Notes Comput. Sci.* **2017**, *10676 LNCS*, 469–474. [Google Scholar] [CrossRef]

25. Han, M.; Tong, M.; Chen, M.; Liu, J.; Liu, C. Application of Ensemble Algorithm in Students' Performance Prediction. In Proceedings of the 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, Japan, 16 November 2017; pp. 735–740. [Google Scholar] [CrossRef]

26. Shanthini, A.; Vinodhini, G.; Chandrasekaran, R. Predicting students' academic performance in the University using meta decision tree classifiers. *J. Comput. Sci.* **2018**, *14*, 654–662. [Google Scholar] [CrossRef]

27. Ma, C.; Yao, B.; Ge, F.; Pan, Y.; Guo, Y. Improving prediction of student performance based on multiple feature selection approaches. In Proceedings of the ICEBT 2017, Toronto, ON, Canada, 10–12 September 2017; pp. 36–41. [Google Scholar] [CrossRef]

28. Tekin, A. Early prediction of students' grade point averages at graduation: A data mining approach [Öğrencinin mezuniyet notunun erken tahmini: Bir veri madenciliği yaklaşidotlessmidotless]. *Egit. Arastirmalari Eurasian J. Educ. Res.* **2014**, 207–226. [Google Scholar] [CrossRef]

29. Pushpa, S.; Manjunath, T.; Mrunal, T.; Singh, A.; Suhas, C. Class result prediction using machine learning. In Proceedings of the 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bengaluru, India, 19 August 2018; pp. 1208–1212. [Google Scholar] [CrossRef]

30. Howard, E.; Meehan, M.; Parnell, A. Contrasting prediction methods for early warning systems at undergraduate level. *Internet High. Educ.* **2018**, *37*, 66–75. [Google Scholar] [CrossRef]

31. Villagrá-Arnedo, C.; Gallego-Duran, F.; Compan-Rosique, P.; Llorens-Largo, F.; Molina-Carmona, R. Predicting academic performance from Behavioural and learning data. *Int. J. Des. Nat. Ecodyn.* **2016**, *11*, 239–249. [Google Scholar] [CrossRef]

32. Sorour, S.; Goda, K.; Mine, T. Estimation of Student Performance by Considering Consecutive Lessons. In Proceedings of the 4th International Congress on Advanced Applied Informatics, Okayama, Japan, 12 June 2016; pp. 121–126. [Google Scholar] [CrossRef]

33. Guo, B.; Zhang, R.; Xu, G.; Shi, C.; Yang, L. Predicting Students Performance in Educational Data Mining. In Proceedings of the 2015 International Symposium on Educational Technology (ISET), Wuhan, China, 24 March 2016; pp. 125–128. [Google Scholar] [CrossRef]

34. Rana, S.; Garg, R. Prediction of students performance of an institute using ClassificationViaClustering and ClassificationViaRegression. *Adv. Intell. Syst. Comput.* **2017**, *50*