# Predicting third party algorithm using data mining

**Mrs.R.Narmatha M.E.,**
Assistant Professor
Department of Computer science and Engineering,
Government College of Engineering-Dharmapuri.
researchnarmatha1988@gmail.com

**T.Vignesh**
Department of Computer Science and Engineering,
Government College of Engineering-Dharmapuri.
vignesht49@gmail.com

**B.Arunkumar**
Department of Computer Science and Engineering,
Government College of Engineering-Dharmapuri.
arunkumarak.ips@gmail.com

**S.Azharuthin**
Department of Computer Science and Engineering,
Government College of Engineering-Dharmapuri.
asarsar097@gmail.com

**ABSTRACT**

**Data mining was designed as the de facto solution to the rising cost of IT storage. With the high cost of data storage devices, as well as the rapid pace at which data is generated, it is costly for businesses or individual users to frequently update their hardware. In addition to reducing storage costs, data outsourcing to the cloud also helps reduce maintenance. The cloud storage shifts the user's data into large data centers that are remotely located and over which the user has no control. This article introduces a data mining application to create student dropout management predictive models. With new records from incoming students, the predictive model can create a short, accurate prediction list that identifies students who need the most support from the student drop-out program. However, this unique feature of the cloud brings many new security challenges that need to be clearly understood and resolved. We provide a schema that provides evidence of data integrity in the cloud that allows customers to verify the accuracy of their data in the cloud. This proof can be agreed by both the cloud and the customer and included in the Service Level Agreement (SLA). The process of encrypting data generally consumes a lot of computing power. In our scheme, the encryption process is very limited to only a fraction of the total data, thus saving the computational time of the client. Data correctness schemes in which a Third Party can audit the data stored in the cloud and assure the customer that the data is safe. The results show that the machine learning algorithm is able to produce an effective predictive model from the existing data of the student's procedural training environment.**

**Keywords: Data mining, Data Integrity, Service Level Agreement, Third Party Audit.**

## 1. INTRODUCTION

Data outsourcing to cloud storage servers increases the trend of many businesses and users due to its economic benefits. In essence, this means that the owner (client) of the data transmits his data to a third-party cloud storage server, which, presumably for a fee, should faithfully store the data and return it to the owner as needed. The storage of user data in the cloud, despite its advantages, holds many interesting security concerns that need to be scrutinized to provide a reliable solution to the problem of local data storage prevention.

Data mining combines machine learning, statistics, and visualization techniques to discover and extract knowledge. Educational Data Mining (EDM) performs tasks such as prediction (classification, regression), clustering, relational degradation (association, correlation, sequential mining and causal data mining), data distillation for human assessment, and model discovery. In addition, EDM can solve many problems based on education. Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information from large amounts of data. It is used to predict the future trends from the knowledge pattern. The main goal of this work is to use data mining methods to find students who are likely to miss the first year of engineering. In this study, the classification task is used to evaluate the dropout data from previous year's students, and since there are many approaches used for data classification, the Bayesian classification method is used here.

Businesses are increasingly turning to more agile IT environments to achieve these goals. One such solution is cloud computing. With cloud computing, tasks can be assigned to a combination of software and services over a network. For example, storing large amounts of data in the cloud reduces costs and maintenance. The customer is not aware of the storage space.

Here, the risk is to change data or manipulate data. Because the customer has no control over data, the cloud provider should assure the customer that data will not be altered. In this paper, we propose a data correctness scheme in which a third party can review data stored in the cloud and assure the customer that the data is secure. This scheme ensures that client-side storage is minimal, which is beneficial for thin clients. A data correcting scheme allows for bandwidth-efficient challenge-response protocols to probabilistically guarantee that a file is available at a remote storage provider.

Knowing the reasons for dropping out can help teachers and administrators take the necessary action to improve their success rate. In order to predict a student's academic achievement, many parameters must be considered. Predictive models, which contain all personal, social, psychological, and other environmental variables, are required to effectively predict students' performance.

## 2. RELATED WORKS

### 2.1 Educational data mining: A review of the state of the art

C. Romero and S. Ventura was said The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice. This process does not differ much from other application areas of DM, like business, genetics, medicine, etc., because it follows the same steps as the general DM process: preprocessing, DM, and post processing. However, it is important to note that in this paper, the term DM is used in a larger sense than the original/traditional DM definition, i.e., we are going to describe not only EDM studies that use typical DM techniques, such as classification, clustering, association-rule mining, sequential mining, text mining, etc., but also describe other approaches, such as regression, correlation, visualization, etc., which are not considered to be DM in a strict sense. Furthermore, some methodological innovations and trends in EDM, such as discovery with models and the integration of psychometric modeling frameworks, are unusual DM categories or are not necessarily seen universally as being DM. From a practical point of view, EDM allows, for example, to discover new knowledge based on students' usage data in order to help to validate/evaluate educational systems, to potentially improve some aspects of the quality of education, and to lay the groundwork for a more effective learning process. Some similar ideas were already successfully applied in e-commerce systems, the first and most popular application of DM, in order to determine clients' interests so as to be able to increase online sales. However, to date, there has been comparatively less progress in this direction in education, although this situation is changing and there is currently an increasing interest in applying DM to the educational environment.

### 2.2 A guide to scaffolding and guided instructional strategies for ITSs

H. K. Holden and A. M. Sinatra was this paper presents a collective student model that has been designed to anticipate the actions that students are likely to take while completing a practical assignment in an educational environment for procedural training. This model is created from activity records or logs collected from students with a similar background that previously completed the same practical assignment. As we will see later, an ITS equipped with this collective student model can use hints to stop students from making certain errors or from floundering with the practical assignment. It is sometimes a good idea to let students make mistakes from which they learn. In other cases, however, it is better to give students the minimum amount of support that they need to progress independently towards problem solving and overcome their zones of proximal development. In this way, each student

learns not from his or her mistakes but with a little bit of help. If necessary, the tutor gradually increases the level of support or scaffolding every time the student makes a mistake or gradually reduces the amount of help provided when the student makes progress. Another reason for helping students not to make mistakes is to prevent student frustration when they fail too often.

## 3. PROPOSED SYSTEM

One of the key concerns that need to be addressed is to provide integrity to the customer; To ensure the accuracy of his data in the cloud. Because the data is physically inaccessible to the user, the cloud should provide the user with an opportunity to verify that the integrity of their data is preserved or compromised. In this article, we introduce a schema that enables data integrity in the cloud, allowing customers to verify the accuracy of their data in the cloud. These data can be agreed by both the cloud and the customer and included in the Service Level Agreement (SLA). It is important to note that our data integrity data only verifies the academic; if the data was illegally modified or deleted. We propose a data correcting scheme that involves encrypting the few bits of data per block of data, thereby reducing the computational burden for the clients. This is based on the fact that a high probability of security can be achieved by encrypting fewer bits rather than encrypting all the data. Client memory overhead is also minimized because it does not store data and reduces bandwidth requirements. In our data integrity protocol, the TPA only needs to store a single cryptographic key, regardless of the size of the data file F and two functions that produce a random sequence. The TPA does not store any data. Before the file is saved to the archive, the TPA prepares the file and appends some metadata to the file and stores it in the archive. At the time of the review, the TPA uses this student's academic data to verify the integrity of the data. It is important to note that our proof of data integrity protocol only verifies the integrity of the data. However, the data can be stored in redundant data centers to prevent data loss due to natural disasters. If the data needs to be modified, which involves updating, inserting, and deleting data on the client side, this requires additional encryption of fewer bits of data. This scheme thus supports the dynamic behavior of data.

## 4. SYSTEM DESIGN

In this scheme, unlike the key hash scheme, only a single key can be used, regardless of the size of the file or the number of files whose unrepeatability is to be checked. In addition, the archive only needs access to a small portion of file F, unlike the keyword scheme, where the archive had to process the entire file F for each protocol verification. The TPA first selects fewer bits of the entire file and prepares the data.

These fewer bits form metadata. This metadata is encrypted and attached to the file and sent to the cloud. Then, whenever the client needs to verify the correctness and availability of the data, it challenges the cloud through TPA and the data it has received are correct, then the integrity is guaranteed. This scheme can be extended for updating, deleting, and pasting data on the client side.

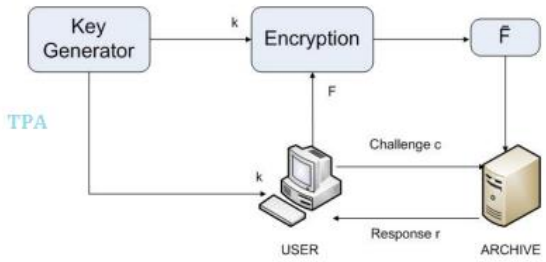



**Figure 4.1: System Design**

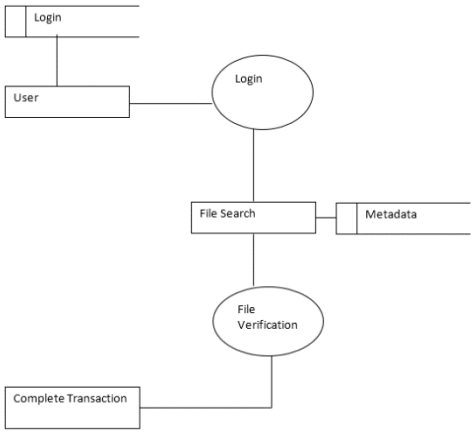



Figure 2. Context level diagram

## 5.METHODOLOGY

### 5.1 Cloud Storage:

Data outsourcing to cloud storage servers increases the trend of many businesses and users due to its economic benefits. Essentially, this means that the owner (client) of the data transmits his data to a cloud storage server of a third party who, presumably for a fee, should faithfully store the data and return it to the owner if necessary.

### 5.2 Simply Archives:

This problem seeks to provide evidence and verify that the data stored by a user in the remote data store in the cloud (called cloud archives or simply archives) is not changed by the archive, thereby ensuring the integrity of the data is. Cloud Archive does not cheat the owner if in this context cheating means that the library can erase some of the data or change some of the data. When developing evidence of ownership of untrusted cloud storage servers, we are often limited by resources at both the cloud server and the client.

### 5.3 Sentinels:

In this scheme, unlike in the key-hash approach scheme, only a single key can be used irrespective of the size of the file or the number of files whose irretrievability it wants to verify. Also the archive needs to access only a small portion of the file F unlike in the key-has scheme which required the archive to process the entire file F for each protocol verification. If the prover has modified or deleted a substantial portion of F, then with high probability it will also have suppressed a number of sentinels.

Encrypting the Meta data: Each of the meta data from the data blocks mi is encrypted by using a suitable algorithm to give a new

modified meta data Mi . Let h be a function which generates a k bit integer αi for each i. This function is a secret and is known only to the TPA.

$$h: i \rightarrow \alpha i,\ \alpha i \in \{0..2n\} \text{----------(2)}$$

For the data (mi) of each data block the number αi is added to get a new k bit number

$$Mi = mi + \alpha i \text{---------- (3)}$$

In this way we get a set of n new data bit blocks. The encryption method can be improved to provide still stronger protection for data

**5.4 Third Party Auditor**

TPA in possession of the public key can act as auditor. It is assumed that TPA is undistorted while the server is untrusted. For application purposes, clients can interact with the cloud servers via CSP to access or retrieve their pre-stored data. More importantly, in practical scenarios, the client can often perform block operations on the data files. The most common forms of these operations are modification, insertion and deletion.

Public audit capability to ensure storage correctness: To allow anyone, not just the clients who originally stored the file on cloud servers, to verify the correctness of the stored data when needed.

Dynamic Data Operations Support: Enables clients to perform block-level operations on the data files while maintaining data integrity. The design should be as efficient as possible to ensure seamless integration of public audit capability and support for dynamic data operations.

**5.5 Verification Phase:**

The verifier prior to saving the file in the archive processes the file and adds some data to the file and stores it in the archive. At the time of verification, the verifier uses this data to verify the integrity of the data. It is important to note that our data integrity protocol only verifies the integrity of data; If the data was illegally modified or deleted. It does not prevent the archive from changing the data.

**6. EXPERIMENTAL RESULTS**

**6.1 Setup phase**

Let the verifier V save the file F with the archive. Let this file F consist of n file blocks. We first process the file and create metadata attached to the file. Each of the n data blocks has m bits in them. A typical data file F that the customer wants to store in the cloud

Generation of Metadata: Let g be a function defined as follows

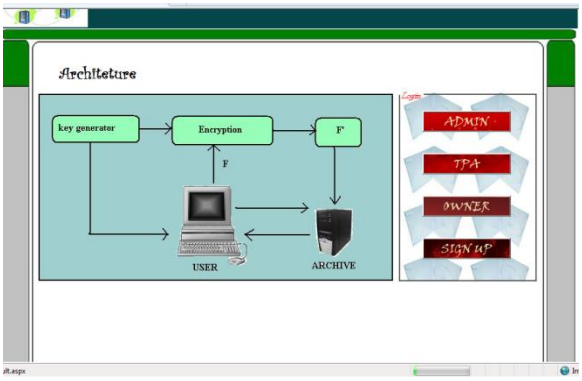
$$g(i, j) \rightarrow \{1..m\},\ i\ \{1..n\},\ j\ \{1..k\} \text{ ----- (1)}$$

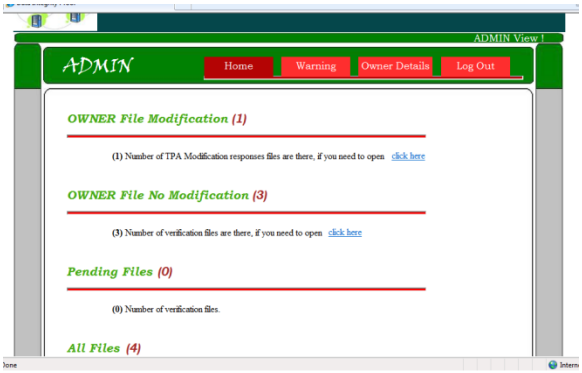
Where k is the number of bits per data block that we want to read as data. The g function generates for each data block a set of k bit positions within the m bits that are in the data block. Therefore, g (i, j) indicates the j-th bit in

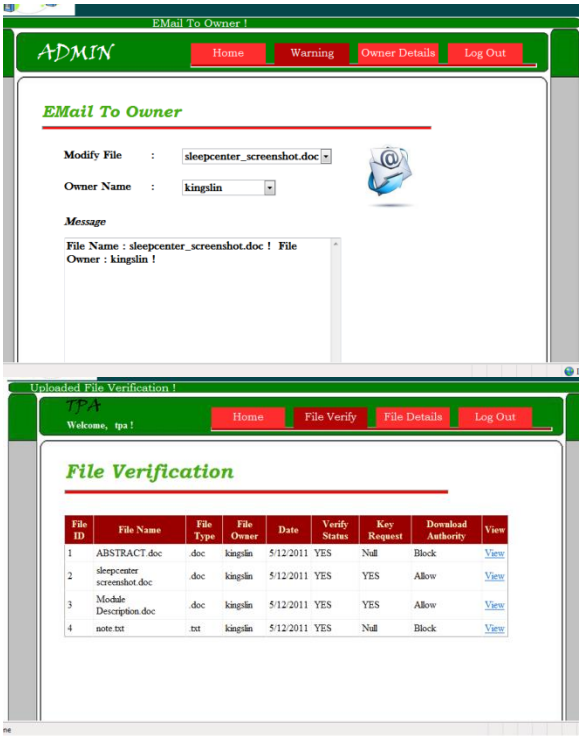
the i-th data block. The value of k lies in the choice of the verifier and is a secret known only to him. Therefore, we get a set of k bits for every data block, and for every n blocks, we get n * k bits. Let mi be the k data bits for the i-th block.

Have the TPA verify the integrity of the F file. She challenges the archive and asks her to answer. The challenge and the answer are compared, and if the result is TRUE, the TPA accepts the integrity proof. Otherwise, if the result of the comparison is FALSE, it rejects the integrity proof. Assuming that the verifier wants to verify the integrity of the i-th block, the TPA prompts the cloud storage server by specifying the block number i and a bit number j using the function g, which only the TPA knows. The TPA also indicates the location where the data corresponding to block i is appended. This data will be a k-bit number. Therefore, the cloud storage server must send the bits for review by the client. The data sent from the cloud is decrypted using the number αi. The corresponding bit in this decrypted data is compared to the bit sent by the cloud. Any discrepancy between the two would lead to a loss of the integrity of the customer's data in cloud storage.

Suppose the verifier wants to check the integrity of the nth block. The verifier challenges the cloud storage server by specifying the block number i and a bit number j generated using the function g, which only the verifier knows. The verifier also specifies the position where the data corresponding to the block i is appended. This data will be a k-bit number. Therefore, the cloud storage server must send k + 1 bits for verification by the client. The data sent from the cloud is calculated using the number .alpha. i is decrypted and the corresponding bit in this decrypted data is compared to the bit sent from the cloud. Any discrepancy between the two would lead to a loss of the integrity of the customer's data in cloud






## 7. CONCLUSION AND FUTURE WORK

The next generation of cloud storage provides a new architecture for storing, managing and analyzing fast-growing machine-generated data. This paper briefly explains the cloud storage, benefits and its features. Our schema is designed to reduce the computing and storage overhead of the Cloud Storage Server. The process of encrypting data generally consumes a lot of computing power. In our scheme, the encryption process is very limited to only a fraction of the total data, thus saving the computational time of the client. Many of the previously suggested schemes require the archive to perform tasks that require a lot of computational power to provide evidence of data integrity. But in our scheme, the archive does not need to fetch and send much data to the client. It uses the linear homomorphism authenticator and random masking to ensure that during the efficient auditing process, the TPA does not learn about the data content stored on the cloud server, not just the burden of the cumbersome cloud user and possibly expensive monitoring eliminates task, but also the fear of the user from outsourced data loss. Therefore, the development will be a challenge for the future. The number of queries that the client can make is also determined a priori. But this number is quite large and may be enough if the data storage duration is short. It will be a challenge to increase the number of queries that use this schema.

## REFERENCE

[1] C. Romero and S.Ventura, “Educational datamining:Asurvey from 1995 to 2005, ” Expert Syst.Appl., vol. 33, no. 1, pp. 135–146, 2007.

[2] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.), vol. 40, no. 6, pp. 601–618, Nov. 2010.

[3] R. S. Baker, “Educational data mining: An advance for intelligent systems in education,” IEEE Intell. Syst., vol. 29, no. 3, pp. 78–82, May/Jun. 2014.

[4] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological

Processes. Cambridge, MA, USA: Harvard Univ. Press, 1978.

[5] A. M. Olney, “Scaffolding made visible,” in Design Recommendations for Intelligent Tutoring Systems. Orlando, FL, USA: U.S. Army Res. Laboratory, 2014, ch. 26, pp. 327–340.

[6] H. K. Holden and A. M. Sinatra, “A guide to scaffolding and guided instructional strategies for ITSs,” in Design Recommendations for Intelligent Tutoring Systems. Orlando, FL, USA: U.S. Army Res. Laboratory, 2014, ch. 22, pp. 265–281.

[7] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker, Handbook of Educational Data Mining. Boca Raton, FL, USA: CRC Press, 2010.

[8] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Za€ıane, “Clustering and sequential pattern mining of online collaborative learning data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 6, pp. 759–772, Jun. 2009.

[9] T. Y. Tang and G. McCalla, “Smart recommendation for an evolving e-learning system: Architecture and experiment,” Int. J. E-Learn., vol. 4, no. 1, pp. 105–129, 2005.

[10] D. Godoy and A. Amandi, “Link recommendation in e-learning systems based on content-based student profiles,” in Handbook of Educational Data Mining. Boca Raton, FL, USA: CRC Press, 2010,
ch. 19, pp. 273–286.

[11] P. Fournier-Viger, R. Nkambou, E. M. Nguifo, A. Mayers, and U. Faghihi, “A multiparadigm intelligent tutoring system for robotic arm training,” IEEE Trans. Learn. Technol., vol. 6, no. 4, pp. 364–377, Oct.–Dec. 2013.

[12] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers “Predicting students drop out: A case study,” in International Working Group on Educational Data Mining. Washington, DC, USA: ERIC, 2009.

[13] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, “Web usage mining for predicting final marks of students that use Moodle courses,” Comput. Appl. Eng. Educ., vol. 21, no. 1, pp. 135–
146, Mar. 2013.

[14] J. A. Lara, D. Lizcano, M. A. Mart_ınez, J. Pazos, and T. Riera, “A system for knowledge discovery in e-learning environments within the European higher education area application to student data from Open University of Madrid, UDIMA,” Comput. Educ., vol. 72, pp. 23–36, Mar. 2014.

[15] C. Antunes, “Acquiring background knowledge for intelligent tutoring systems,” in Proc. 2nd Int. Conf. Educ. Data Mining, 2008, pp. 18–27.

[16] A. Hershkovitz and R. Nachmias, “Learning about online learning processes and students’ motivation through Web usage mining,” Interdisciplinary J. Knowl. Learn. Objects, vol. 5, pp. 197–214, 2009.

[17] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, “Emotion sensors go to school,” in Proc. Conf.

Artif. Intell. Educ.: Building Learn. Syst. That Care: From Knowl.
Representation Affective Model., 2009, pp. 17–24.
[18] T. Barnes and J. Stamper, "Automatic hint generation for logic proof tutoring using historical data," J. Educ. Technol. Soc., vol. 13, no. 1, pp. 3–12, 2010.
[19] M. Mavrikis, "Modelling student interactions in intelligent learning environments: Constructing Bayesian networks from data," Int. J. Artif. Intell. Tools, vol. 19, no. 6, pp. 733–753, Dec. 2010.
[20] K. Porayska-Pomsta ,M. Mavrikis, S. D'Mello, C. Conati, and R. S. J. d. Baker, "Knowledge elicitation methods for affect modelling in education," Int. J.Artif. Intell. Educ., vol. 22, no. 3, pp. 107–140, 2013.
[21] A. Bogar_ın, C. Romero, R. Cerezo, and M. S_anchez-Santill_an, "Clustering for improving educational process mining," in Proc. 4th Int. Conf. Learn. Anal. Knowl., 2014, pp. 11–15.
[22] A. Pe~na-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," Expert Syst. Appl., vol. 41, no. 4, pp. 1432–1462, Mar. 2014.
[23] K. Sukhija, M. Jindal, and N. Aggarwal, "The recent state of educational data mining: A survey and future visions," in Proc. IEEE 3rd Int. Conf. MOOCs Innovation Technol. Educ., 2015, pp. 354–359.
[24] L. Razzaq, et al., "The assistment project: Blending assessment and assisting," in Proc. 12th Annu. Conf. Artif. Intell. Educ., 2005, pp. 555–562.
[25] S. Ritter, J. Anderson, K. Koedinger, and A. Corbett, "Cognitive tutor: Applied research in mathematics education," Psychonomic Bulletin Rev., vol. 14, no. 2, pp. 249–255, 2007.
[26] S. Ritter, R. Carlson, M. Sandbothe, and S. E. Fancsali, "Carnegie learning's adaptive learning products," presented at the 8th Int. Conf. Educ. Data Mining, Madrid, Spain, 2015.
[27] F. Vicente, S. Adjei, T. Colombo, and N. Heffernan, "Building models to predict hint-or-attempt actions of students," presented at the 8th Int. Conf. Educ. Data Mining, Madrid, Spain, 2015.
[28] R. S. Baker, et al., "Towards sensor-free affect detection in cognitive tutor algebra," in Proc. 5th Int. Educ. Data Mining Soc., 2012, pp. 126–33.