DETECTING THE AFFECTED FILES IN CLOUD BASED ON ALGORITHM NAME

[1]Mrs. K. Mahalakshmi, [2]M. Lakshmi,
[1]Assistant Professor, [2]UG Reasearch Scholar
[1]Department of B. Com (Business Analytics)
[1]PSGR krishnammal college for women,Coimbatore,Tamilnadu,India.

***ABSTRACT*:**

This project aims to analyze the cloud computing dataset and find the affected files and the secured files using the algorithm name and give the result on highly secured algorithm. The resulting design is able to facilitate efficient server-side ranking without losing keyword privacy. This will help the user to know the secrecy of their data. This paper also analyzes each algorithm and its level of secrecy.

Keyword: *cloud computing, Random forest classifier, data analysis.*

I. INTRODUCTION

Data analysis is the process of inspecting, transforming, analyzing data sets to get the insights from the data to take business decisions using the machine learning algorithm. For that we have used 'Jupyter notebook'. The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. Here, Random Forest algorithm is used to analyze the cloud computing dataset. Cloud computing is dramatically changing the way that organizations manage their data, owing to its attractive features such as robustness, low cost, and ubiquitous nature. However, privacy concerns arise whenever sensitive data is outsourced to the cloud where the data is processed and stored. As the data, in most cases encrypted, have to be not only stored, but also processed in clouds, the cryptography-based data confidentiality and integrity protection approaches are not adequate to satisfy the security requirements.Privacy preserving in cloud environments includes two aspects: data processing security and data storage security. Data processing security covers the issues of how to protect user privacy at runtime in a virtualized cloud platform. Data storage security covers the issues of guaranteeing user data privacy when the data is stored in data center. Privacy is an important issue for cloud computing, both in terms of legal compliance and user trust, and needs to be considered at every phase of design.

II. OBJECTIVES

The objective of the paper is to analyze the affected files in cloud based on the algorithm name by using the Random Forest Classifier algorithm and identify which algorithm is highly secured and which algorithm is highly affected.

III. RELATED WORK

Cloud computing offers the prospect of on-demand, elastic computing, provided as a utility service, and it's revolutionizing many domains of computing. The shift in paradigm that accompanies the adoption of cloud computing is increasingly giving rise to security and privacy considerations concerning facets of cloud computing like multi-tenancy, trust, loss of control and accountability. Consequently, cloud platforms that handle sensitive information are required to deploy technical measures and organizational safeguards to avoid data protection breakdowns which may end in enormous and dear damages. Hence, with the expansion of cloud computing in recent times, privacy and data protection requirements are evolving to guard individuals against surveillance and data disclosure. Some samples of such protective legislation are the EU Data Protection Directive (DPD) and therefore the US insurance Portability and Accountability Act (HIPAA), both of which demand privacy preservation for handling personally identifiable information. This dissertation focuses on the planning and development of several systems and methodologies for handling sensitive data appropriately in cloud computing environments. The key idea behind the proposed solutions is enforcing the privacy requirements mandated by existing legislation that aims to protect the privacy of individuals in cloud-computing platforms. [2]

Data security has consistently been a serious issue in information technology. In the cloud computing environment, it becomes particularly serious because the info is found in several places. Data security and privacy protection are the 2 main factors of user's concerns about the cloud technology. Though many techniques on the topics in cloud computing are investigated in both academics and industries, data security and privacy protection are getting more important for the longer-term development of cloud computing technology. Data security and privacy protection issues are applicable to both hardware and software in the cloud architecture.[5]

Over the past few years, major IT vendors (such as Amazon, Microsoft and Google) have provided virtual machines (VMs), via their clouds, that customers could rent and utilize hardware resources and support live migration of VMs additionally to dynamic load-balancing and on-demand provisioning. This means that, by renting VMs via a cloud, the whole datacenter footprint of a contemporary enterprise is often reduced from thousands of physical servers to a couple of hundred (or even just dozens) of hosts. While it's practical and price effective to use cloud computing during this way, there are often issues with security when using systems that aren't provided in-house. To look into these and find appropriate solutions, there are several key concepts and technologies that are widely utilized in cloud computing that require to be understood, such as virtualization mechanisms, varieties of cloud services, and “container” technologies.[1]

The recent popularity of cloud computing, data owners now have the chance to outsource not only their data but also processing functionalities to the cloud. Because of data security and private privacy concerns, sensitive data (e.g., medical records) should be encrypted before being outsourced to a cloud, and therefore the cloud should perform query processing tasks on the encrypted data only. These tasks are described as Privacy-Preserving Query Processing (PPQP) over encrypted data. Based on the concept of Secure Multiparty Computation (SMC), SMC-based distributed protocols were developed to permit the cloud to perform queries directly over encrypted data. Several queries were considered in an effort to make a well-defined scope. It includes the k-Nearest Neighbor (KNN) query, advanced analytical query, and correlated range query. The proposed protocols utilize an additive homomorphic cryptosystem and/or a garbled circuit technique at different stages of query processing to achieve the best performance.[3]

Cloud data security may be a major concern for the cloud user while using the cloud services provided by the service provider. To ensure correctness of data here propose the task of allowing a third-party auditor. On behalf of cloud user request to verify the integrity of knowledge stored within the cloud is completed by TPA. The advantage of TPA is that there's no additional online burden to user. [4]

IV. METHODOLOGY

Random Forest Algorithm

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

1) Pick N random records from the dataset.
2) Build a decision tree based on these N records.
3) Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
4) In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

Random Forest Regression

In this section we will study how random forests can be used to solve regression problems using Scikit-Learn. In the next section we will solve classification problem via random forests.

Random Forest Classifier

Problem Definition

The task here is to predict whether a bank currency note is authentic or not based on four attributes i.e. variance of the image wavelet transformed image, skewness, entropy, and curtosis of the image.

s.no	Field Name	Data Type	Description
1	id	int64	0-3000 id’s
2	Is Encrypted	Object	Encrypted are not
3	Algorithm Name	Object	Different types of algorithm used
4	File Protected	Object	File is protected or not

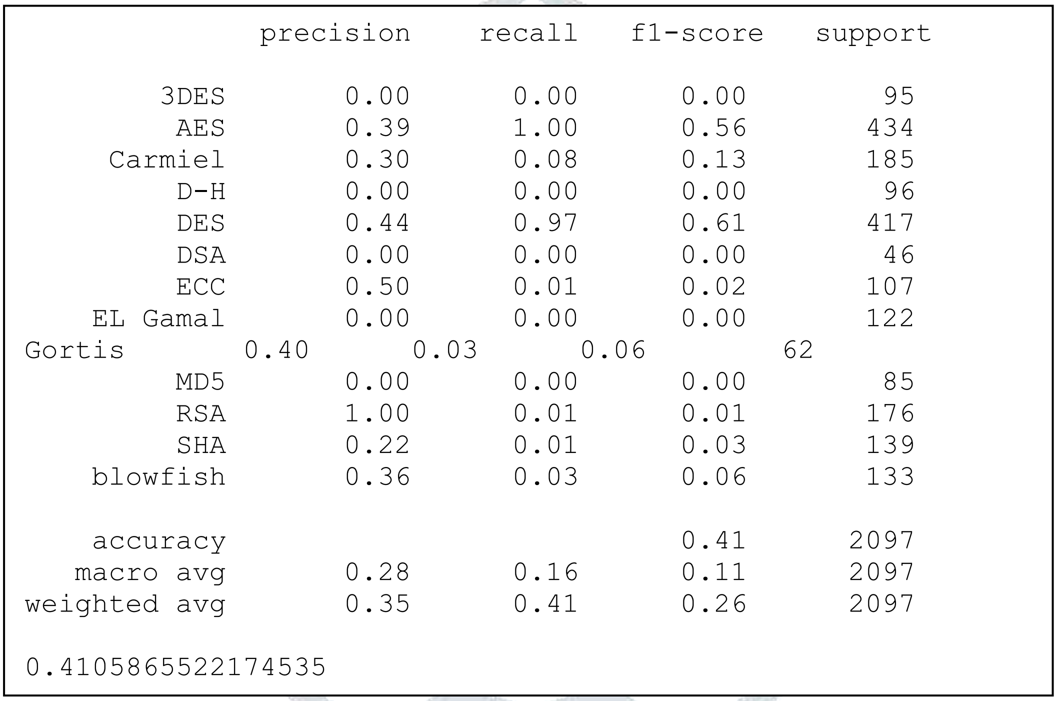
5	Is file attacked	Object	Files are attacked or not
6	File attacked method	Object	Different attacking method
7	TPA verified	Object	TPA verified or not
8	Frequency of file attack	int64	Attacked files frequency

Table 1: Selected variables from cloud computing record.

The above table displays the attributes which are used in the data with description and details about the value in that.

Using this, the files were analyzed to find which algorithm is highly secured.

V. RESULT



```
              precision    recall  f1-score   support

        3DES       0.00      0.00      0.00        95
         AES       0.39      1.00      0.56       434
     Carmiel       0.30      0.08      0.13       185
         D-H       0.00      0.00      0.00        96
         DES       0.44      0.97      0.61       417
         DSA       0.00      0.00      0.00        46
         ECC       0.50      0.01      0.02       107
    EL Gamal       0.00      0.00      0.00       122
Gortis          0.40      0.03      0.06          62
         MD5       0.00      0.00      0.00        85
         RSA       1.00      0.01      0.01       176
         SHA       0.22      0.01      0.03       139
    blowfish       0.36      0.03      0.06       133

    accuracy                           0.41      2097
   macro avg       0.28      0.16      0.11      2097
weighted avg       0.35      0.41      0.26      2097

0.4105865522174535
```

Table 2

The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

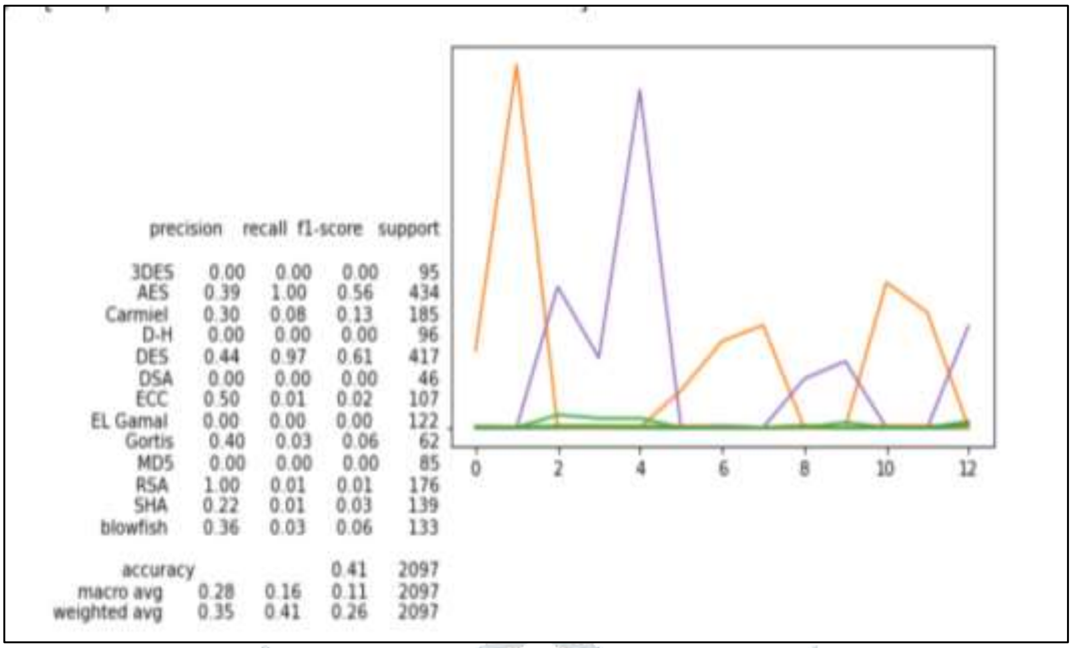
The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

When true positive + false positive == 0, precision is undefined. When true positive + false negative == 0, recall is undefined. In such cases, by default the metric will be set to 0, as will f-score, and Undefined Metri Warning will be raised. This behavior can be modified with zero_division.

F1 score is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. Precision is also known as positive predictive value, and recall is also known as sensitivity in diagnostic binary classification.

The **F_1** score is the harmonic mean of the precision and recall.

The support is the number of occurrences of each class in y_true. And table 2 describe the precision, recall and F1 Score and support for the attribute Algorithm name and the accuracy value in percentage.



```
              precision    recall  f1-score   support

        3DES       0.00      0.00      0.00        95
         AES       0.39      1.00      0.56       434
     Carmiel       0.30      0.08      0.13       185
         D-H       0.00      0.00      0.00        96
         DES       0.44      0.97      0.61       417
         DSA       0.00      0.00      0.00        46
         ECC       0.50      0.01      0.02       107
    EL Gamal       0.00      0.00      0.00       122
      Gortis       0.40      0.03      0.06        62
         MD5       0.00      0.00      0.00        85
         RSA       1.00      0.01      0.01       176
         SHA       0.22      0.01      0.03       139
    blowfish       0.36      0.03      0.06       133

    accuracy                           0.41      2097
   macro avg       0.28      0.16      0.11      2097
weighted avg       0.35      0.41      0.26      2097
```

Figure.1

The above figure-1 represents the accuracy achieved for by our random forest classifier with 20 trees is 41%.41% is an average accuracy, so there is much point in increasing our number of estimators.

We can see that increasing the number of estimators is further improve the accuracy.To improve the accuracy, we would suggest you to play around with other parameters of the Random forest classifier class and see if you can improve on our results. AES algorithm is highly secured algorithm because it has high supported value(434).

VI. CONCLUTION

In this paper, jupyter is used to analyze the cloud computing data using the Random Forest algorithm. The main intention of this paper is to help the user to know that their files are secured or not. Here we have used 3000 records of data which contain various attributes. Using these data were analyzed and provided necessary data.

References:

1. Ali Gholami and Erwin Laure,” SECURITY AND PRIVACY OF SENSITIVE DATA IN CLOUD COMPUTING: A SURVEY OF RECENT DEVELOPMENTS”, HPCViz Dept., KTH- Royal Institute of Technology, Stockholm, Sweden

2. Ali Gholami, “Security and Privacy of Sensitive Data in Cloud Computing”, Doctoral Thesis Stockholm, Sweden 2016

3. Elmehdwi, Yousef M., "Privacy-preserving query processing over encrypted data in cloud" (2015). Doctoral Dissertations. 2442.
4. Miss. PratikshaMeshram, Prof. RoshaniTalmale, Prof. G. Rajeshbabu, A System of Privacy Preserving Public Auditing for Secure Cloud Storage System, IJERT, Volume 03, Issue 08 (August 2014)
5. Yunchuan Sun[1], Junsheng Zhang, [2], Yongping Xiong[3], Guangyu Zhu[4]Data Security and Privacy in Cloud Computing, Article first published online: July 16, 2014; Issue published: July 1, 2014 Received: April 25, 2014; Accepted: June 26, 2014, China
6. Leo Breiman Statistics Department University of California Berkeley, CA 94720 January 2001.
7. Qiong Ren a), Hui Cheng b) and Hai Han School of Mathematics and Computer Science, Jianghan University, Wuhan, China.
8. 1Vrushali Y Kulkarni, 2 Pradeep K Sinha International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 11, May 2014
9. Amit, Y. & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*, 1545–1588.
10. Amit, Y., Blanchard, G., & Wilder, K. (1999). Multiple randomized classifiers: MRCL Technical Report, Department of Statistics, University of Chicago.