Missing Values Prediction for Cyber Vulnerability Analysis in Academic Institutions

Bhavya Agrawal
University School of Information, Communication and Technology Guru Gobind Singh Indraprastha University, Delhi, India

Anurag Jain
University School of Information, Communication and Technology Guru Gobind Singh Indraprastha University, Delhi, India

ABSTRACT

In this paper, a survey-based study has been done to analyze the cyber security vulnerability of higher education institutions to identify the areas that are more prone to cyber threats at different user levels (System Administrator and Students & Faculty). One of the major elements of data mining- prediction of Missing Values has been amalgamated with vulnerability analysis of academic institutes to improve their practices and compliance of information security. These predictions help in identifying associations and handling missing data due to lack of awareness among users for more effective vulnerability analysis of the cyber security in academic environments. Subsequently, it will lead to formation of essential security guidelines that institutes can adopt to avoid above mentioned risks. Two theories have been proposed to identify the cyber vulnerabilities based on Questionnaire filled by different user levels. Prediction of missing values has also been evaluated after pre-processing and tried to filled the blank entities in the Questionnaire. The result shows that, after the prediction of missing values there is still significant number of students and faculty who are confused about the HR Policies of their institutes making their information security vulnerable. Hence guidelines to mitigate vulnerability issues have been proposed in this research work.

General Terms

Cyber Security, Vulnerability Analysis, Naïve Bayes Classification, Data Mining Prediction, Higher education institutes

Keywords

Cyber security, Vulnerability Analysis, Information Security, Security Guidelines, Academic Institutes, Naïve Bayes algorithm, Prediction

1. INTRODUCTION

The educational institutes are one such domain where cyber criminals are gaining access very swiftly due to increase of Internet Services, lack of awareness among students, lack of training of faculties, noncompliance of security standards and policies or lack of security guidelines. The sensitive and confidential data of the institutes faces the danger of breach and loss due to lack of awareness of policies and their compliance. Therefore, cyber-attacks are rising rapidly at an alarming rate in these academic institutions. The information or data of the staff and the institute is always at risk of unauthorized access. This exposes them to security breaches, network outrages, unavailability of information. The educational data is an irreplaceable asset and very confidential in nature hence it is the ultimate responsibility of the management to ensure that this data is adequately protected. If it fails to protect the integrity and security of such data than it welcomes a host of potential problems ranging from the charges of remissness and incompetence, to law suits charging "computer malpractice,"[1]. Few instances of cyber security breach have happened in various academic institutes recently. The University of Maryland fall prey to the cyber security attack that exposed personally identifiable information (PII) records in February 18, 2014. The North Dakota University System announced that a server was hacked by some unauthorized entities a week after the incident of Indiana University which resulted in leakage of data containing names and SSN Numbers of approximately 290,000 students and about around 780 faculty.[2].There are various cyber security issues or challenges that are faced by academic institutes. Bob Turner, CISO at University of Wisconsin-Madison stressed on the fact that cyber security awareness or proper security training takes a backseat due to the long working hours of faculty, high load of academic courses on students. Improper implementation of the policies laid for security in institutes also makes the system weak and vulnerable the dearth of proper backup and recovery techniques can also be cited as the major area of concern. Cloud Computing offers great technology for handling data of universities but brings along some major securities issues due to its open nature [3]. These are some of the identified loopholes that can put the cyber security of academic institutes at risk. Due to these vulnerabilities the cyber criminals try to steal, destroy the information in academic institutes and mint out advantage from it using their fraudulent schemes. These repositories form appealing targets for the attackers. Although universities are now getting more aware of the challenges of security and are adopting various measures and policies but there is gap in terms of their implementation and execution at various levels. Thus, there is a need to form strong security policies framework in the institutes and their 100 percent compliance with proper knowledge transmission and trainings and various other methods. This research aims to do efficient analysis of cyber security vulnerabilities and identified cyber issues, using a survey-based approach in academic institutes. This survey has covered over 9 important categories of cyber security at various levels of institute to find out the gap between their knowledge. A questionnaire was structured to be filled by students, faculties and system administrators. The data encountered many missing responses which may be due to the reason that the students are less aware of the distinct categories of cyber *security*, the purpose of the usage of different tools, lack of knowledge of security policies and their implementation in their institute. Due to the presence of huge amount of missing values this research has lot of scope to handle these values by prediction and find some associations between some policies. These predicted values will lead to better analysis and will further help in carving out an effective set of security guidelines and policies that can be used by various academic institutes to keep a check on existing cyber security vulnerabilities**.** These predictions are also useful in identifying new associations in the data which could further help in framing more compact and effective questionnaire**.** Thus, this paper aims to design a model to predict whether personal information of staff and students is secured or not based on association whether the information security training has been provided to students or not in their respective institutes using Naïve Bayes classification technique and further analyze the vulnerabilities of security policies in academic institutes.

The rest of the research paper is as follows: Section II discusses about comprehensive related work, Section III gives the overview of the research methodology adopted for the analysis, Section IV describe the implementations details, Section V and VI presents the results of the prediction and the vulnerability analysis and Section VII gives the conclusion of the paper and mentions the Future Works.

2. RELATED WORK

Cyber security has become a very alarming and rapidly increasing issue for the internet users all across the world. The Indian users are no exception to it and are also becoming its victim. Cyber Security has become a necessity as today in the world of web, data has become the most important entity to be used. It seeks its usage in every domain such as government, medical, banks, academic institutions etc. Thus, the cybersecurity of such digital data is a very important topic in this era. The research happening in this field is tremendous at both national and international level.

Narendra Modi, PM of India has laid stress on the fact that the cybersecurity threats should be dealt with the highest priority among all the national issues [4]. Very few researches have been done in the field of cyber security for academic institutions in India and abroad and need of proper security mechanism persists which should be addressed vigilantly. Research demonstrates that the academic institutions, especially universities, have become the new hot targets for cyber criminals due to two main reasons. The first cause is high computing power of data and the second is direct access to the public and its constituencies [5]. In literature, a gap analysis has been presented on the prevailing vulnerabilities in information security policies of Indian academic institutes with that of western institutes [6]. An investigation has been done to identify the recent security issues and major areas of concern to ensure more secured campuses. Another area which has been examined is the network usage security policies for academic institutions and the authors have recommended some policies guidelines that will help institutes to control network security efficiently [7]. A white paper from SANS, has presented, defined, examined compliance and suggested about the various policies for cybersecurity in academic institutions. [8]. Another white paper from London school has discussed policies based on access control that should be adopted by the educational institutions to give the access privileges of various systems to authorized people only [9]. This paper has described that education institutions are the treasure chests for criminals of cyber due to massive amount of data they handle, and the increase usage of cloud computing resources also gives way to attackers to target them due to its weak security [10].In this SITS approach was used as a technique in the higher education information security, although the technique was simple and easy to use but due to departmental and academic diversities other approach was used that is RITSB[11]. The importance of antecedent and measure in effecting the awareness of information security of the user has been discussed in this paper and concluded the religious indicator and training program factor based on user perspective as the most important to increase ISA in higher education [12] [13]. This paper identified some already existed policies regarding cybersecurity, tools and technology used by students and provided guidelines for improving data security-awareness at higher educational institutions [14]. Neo-Institutional Theory (NIT) stated that the factors such as regulatory and social normative pressures are more impactful for conformance of security policies in higher institutions [15]. This journal compared the various classification techniques such as to handle missing values and concluded hybrid approach gave better graceful results [16]. Bayesian classification method on some student database of some colleges collected through a questionnaire and college database to predict the student division based on previous year database has been used [17]. The paper has done comparative discussion on classifiers Naïve Bayes and J48 on the dataset of a bank with the intent to increase accuracy of the finding the defaulters with the help of Weka tool [18].

3. RESEARCH OVERVIEW

In this project, a survey has been conducted among some academic institutes by preparing a questionnaire to investigate the level of awareness among different users about the security policies in the institutes. The questions have been selected after a thorough brainstorming session and discussions with cyber Security professionals and studying various research papers related to cyber security. This questionnaire covers the 4 essential areas of cyber security to develop an initial phase security model.

3.1 Cyber Security Areas

The following are the 4 essential areas of cyber security that are covered in this research.

3.1.1 HR Security

It includes security of the personal information and their authorization against unwanted access. It lays stress on the security the data at various user levels. Eg. Does your institution secure personal information of students, faculty?

3.1.2 Risk Management

It involves management of the unknown risks that might happen. It checks whether proper backups are taken or not regular intervals. It also lays down other plans in case of disaster and their recovery. Eg. Is your data backup process frequency consistent?

3.1.3 Network Security

It handles the security of the various networks within an institute or outside the institute. Network Outrages, connectivity issues, virus attacks in Lans and Wifi network within the institutes all comes under network security issues. The layers of security used within the wifi networks. Eg Does your institution have internet connectivity at following places at labs, hostels etc?

3.1.4 Acquisition and maintenance

It ensures the security of the acquired assets and their maintenance by following various guidelines which may or may not be vendor specific. The various coding processes followed whiling including and acquiring he data, validation of the data. The ultimate goal is to practice proper process and rules while gathering and processing the data in institutions.Eg Does your institution have process for validating the security of purchased software products and services?

There are three main types of valid responses in the questionnaire.

1. Y-to indicate Yes and show that the user is very sure that the answer is positive.
2. N-to indicate No and to show that the user is very sure that the answer is negative.
3. C-to indicate Can't Say response and to show that user is not sure or confused about the Survey question.

However, we came across many missing responses as well. Our aim will be to predict them appropriately and do the further analysis.

3.2 Survey Target Group

The target group of this questionnaire are the system administrators, faculty and the students are the key users of any academic institutes.

3.2.1 Student Domain

This domain includes the students of the academic institute. They are the most vulnerable users due to the highest level of lack of awareness of security policies or cybersecurity issue amongst them.

3.2.2 Faculty Domain

Faculty domain includes all faculty members of the academic institutes.

3.2.3 System Administrator

This Domain includes the administrative staff which are responsible to deal and manage systems that handle confidential, sensitive, personal data. They are the ones who manage the websites of institutions and networks within the institutions. They in fact have in-depth knowledge of the security policies and also play an important in managing institute web systems. Usually includes the security and network administrator of the academic institutes.

3.3 Dataset

This questionnaire covers the 4 essential areas of Cyber Security as mentioned earlier and filled by the system administrators, faculties and students of the various institutes. Before presenting this questionnaire in front of the cyber users of the institute information have been provided about cyber vulnerabilities and terminologies so that the users could get knowledge about the questionnaire.

The responses are stored into the database which are further used for the vulnerability analysis of the academic institutions.

Currently our database contains data of 6 colleges in Delhi region area of total 237 records having 7110 values for 30 primary attributes. It includes students, faculty, and system administrator data.

3.4 Naïve Bayes Classification Technique

Bayesian classifiers are a type of probabilistic classifier that predicts whether a particular record belongs to a particular class or not [20]. Bayes' Theorem is its foundation. It assumes and works on the theory that the probability of an attribute to belong to a particular given class or group is not dependent or rather independent of the values of other attributes present in the data. This research includes use of this classification technique. Bayes rule is a method to approximate the probability of a property if the dataset is given as evidence [17].Bayes theorem is formulated as

$$P(w_i|x_i) = P(x_i|w_i)P(w_i)/(P(x_i|w_i) + P(x_i|w_2)P(w_2))$$

Given a training set the naïve Bayes algorithm first calculates the prior probability $P(c_j)$ for every class by counting how often each class occurs in the training data. Then all the probabilities of every attribute value are calculated by $P(x_i)$. After that the probability $P(x_i|c_j)$ is calculated by summing how often every value occurs in the class in the training data. When grouping the target label, the conditional and prior probabilities calculated from the training set are used to make the prediction. Then find $P(t_i|c_j)$ by following

$$P(t_i|c_j) = \prod_{k=1}^{p} P(x_{ij}|c_j)$$

To calculate $P(t_i)$ we can estimate the probability that ti is in each class. The probability that t_i is in a particular class is the product of the conditional probabilities for each attribute value. The class with the maximum probability is the final class chosen for the tuple [21]. These naïve Bayes approach brings along with itself various advantages like ease of use, just one scan of training data is needed, does not require large amount of data for classification [22].

3.5 Research Methodology

The Figure 1 demonstrates the operational flow process to identify Cyber vulnerability at Academic level. This process consists of 6 iterative phases and it is a continuous process to update the questionnaire. The updation of questionnaire is based on the feedback given by the various users and after identifying the missing values in the dataset.

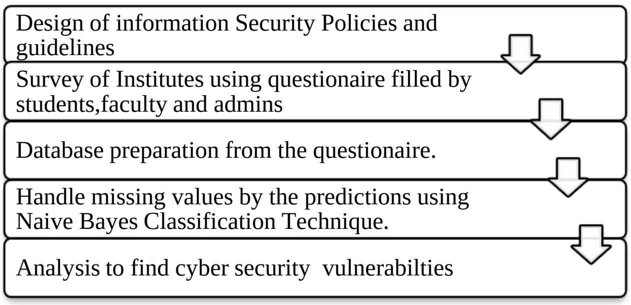
Design of information Security Policies and guidelines
Survey of Institutes using questionaire filled by students,faculty and admins
Database preparation from the questionaire.
Handle missing values by the predictions using Naive Bayes Classification Technique.
Analysis to find cyber security vulnerabilties

Figure 1. Operational Flow Process to identify Cyber Vulnerability at Academic level

The main problem statement that has been considered is prediction of the missing responses using Naïve Bayes classification technique. To carry out the experimental research, few assumptions have been made. These assumptions are required to understand the knowledge of participants and to predict the missing value analysis-

1. In some questions (as stated in Table 1) that has been chosen to be predicted, the respondents have answered the details of how the data is secured. So, it has been assumed that if they have some details then they know the security is provided so the response assumed is Yes.
2. Missing responses of survey questionnaire (e.g Q2 TO Q7 as given in Table 1) are filtered out during the prediction.
3. As all the attributes might not be effective in predicting whether security is provided to personal information. So, the other questions from the same category has been chosen for prediction and also find some association among them. Following Table 1. shows the instances of sample questions used from the questionnaire for the prediction model.

Table 1. Set of Questions used in the prediction model

Sno	Questions
Q1	Does your institution secure personal information of students, faculty members, administration?
Q2	Does your institution provide information security awareness training to the staff?
Q3	Does your institution conduct cyber training for the students?
Q4	Does your institution provide information security awareness training to the students?
Q5	Does your institution conduct cyber training for the faculty members?
Q6	Does your institution conduct cyber training for the non-teaching staff?
Q7	Does your institution conduct cyber training for system administrators?

4. IMPLEMENTATION

After the preparation of database, the model is proposed to predict the weaknesses of Academic institutions using RapidMiner Tool. It is a java based open source tool which was developed by the company Rapidminers. The RapidMiner Marketplace offers a great environment and platform for developers to develop data analysis techniques and share them to the community [19]

4.1 Vulnerability Prediction Model

There are 4 phases of Prediction Model as shown in Figure 2.

1. In Phase 1 the data has been pre-processed using Multiply, FilterExample, Select Attribute and Set Role Operator.
2. In Phase 2 the pre- processed data is Trained using Naïve Bayes Classification Technique.
3. In Phase 3 Validation of the data has been performed using Split Validation Operator, Performance Evaluation Operator by taking 65% of training and 35% testing data.
4. In Phase 4 Prediction of missing responses is done using Multiply, FilterExample, Select Attribute, and Apply Model and Operator.

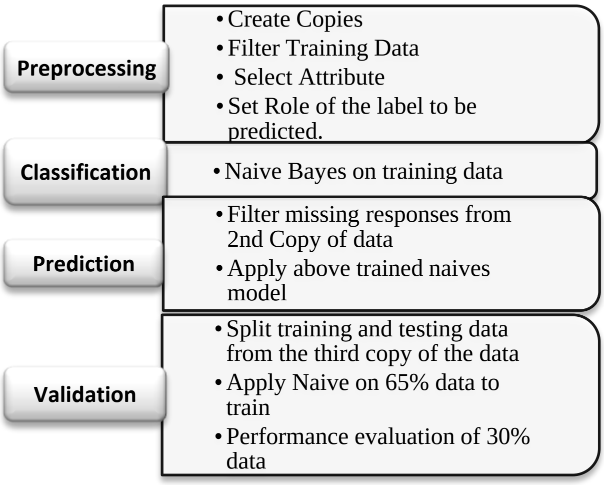
Preprocessing
• Create Copies
• Filter Training Data
• Select Attribute
• Set Role of the label to be predicted.
Classification
• Naive Bayes on training data
Prediction
• Filter missing responses from 2nd Copy of data
• Apply above trained naives model
Validation
• Split training and testing data from the third copy of the data
• Apply Naive on 65% data to train
• Performance evaluation of 30% data

Figure 2. Phases of Prediction Model

After the prediction model, vulnerability analysis has been performed by comparing the responses of system administrator with the responses of students and faculties.

4.2 Vulnerability Analysis

Cyber security vulnerability of academic institutes is being analyzed. Two theories have been used to calculate vulnerability.

4.2.1 Theory 1

Vulnerability has been measured by finding the maximum % of responses for the various questions under different categories selected. If maximum % of Responses is C or Can't Say then it can be concluded as vulnerable. Vulnerability % will be the % of Can't Say Responses i.e C%.

4.2.2 Theory 2

Other Way to detect vulnerability used is by comparing the maximum response given by Admin and maximum response given by students and faculties for some question. If it comes out to be different then it is vulnerable and if it comes out to be same but percentage of maximum responses by students and faculties is less than threshold value (say 50 %) then also it is considered vulnerable else it is not vulnerable. Formula for vulnerability % is mentioned in Table 2 as below. The abbreviations used in Table 2 are explained in Table 3.

Table 2. Pseudo Code for Vulnerability Analysis of Questions.

Theory	Pseudo Code
Theory 1	*While Question in Qn* *Calculate Max Response ← Max (C%, Y%, N%)* *If Max Response=C% then Qn←Vulnerable* *If Qn='Vulnerable' then Vulnerability %←C%*

Theory 2	*while Question in Qn*
	Calculate Max A Response ← Max (C% A, Y% A, N% A, Missing % A) where Question=Qn
	Calculate Max FS Response ← Max (C% FS, Y% FS, N% FS, Missing % FS) where Question=Qn
	if [Max A Response]! = [Max FS response] or ([Max A Response] = [Max FS response] and [MaxFS %] <⍰) or ([Max A Response] = [Max FS response] and [Maximum A %]- [Maximum FS %]>= ⍰) then Qn←"Vulnerable" else Qn← "Not Vulnerable"
	if Qn = 'Vulnerable' and [Max A Response]! = [Max FS response]
	then Vulnerability % ←[Maximum FS %]
	else if Qn='Vulnerable' and (([Max A Response] = [Max FS response] and [Maximum FS %] < ⍰) or ([Max A Response] = [Max FS response] and [Maximum A %]- [Maximum FS %]>= ⍰))
	then Vulnerability % ←1- [Maximum FS %]

Table 3. List of Abbreviations used in Table 2

Symbols	Description
Y%	Y% is number of records where response is 'Y' or Yes for RecordType in Admin, Student, Faculty
N%	N% is number of records where response is 'N' or No for RecordType in Admin, Student, faculty
C%	C% is number of records where response is 'C' or Can't Say for RecordType in Admin, student, faculty
Max Response	Maximum % out of Y%, N%,C%
Y% A	% of records with Response as "Y" and RecordType as Admin.
N% A	% of records with Response as "N" and RecordType as Admin
C% A	% of records with Response as "C" and RecordType as Admin
Missing % A	% of records which are Missing and RecordType as Admin
Y% FS	% of records with Response as "Y" and RecordType as Student and Faculty.
N% FS	% of records with Response as "N" and RecordType as Student and Faculty.
C% FS	% of records with Response as "C" and RecordType as Student and Faculty.
Missing % FS	% of records which are Missing and RecordType as Student and Faculty.
Maximum FS %	Maximum % of Y % FS, C% FS, N% FS,Missing %FS
Maximum A %	Maximum % of Y % A, C% A, N% A,Missing %A
Max A Response	Maximum Response of Faculty and students
Max FS Response	Maximum Response of Admin
Vulnerabilty %	Vulnerability % is the percentage with which question is Vulnerable.
Qn	Qn is the question whose responses are to be compared where n represents the sequence number
µ	Threshold Value

5. EXPERIMENTAL RESULTS

The analysis of number of missing responses is done for questions in Table 1 following which the prediction model has been developed to predict the missing responses. Subsequently the vulnerability analysis of the data has been done at different levels of academic institutes. The results have been tabulated and demonstrated in the following tables 5 to 7(a-b) and figure 3 to 9(a-b).

5.1 Missing Response Analysis

The category that has been chosen for the analysis is HR Security category. The Figure 3 illustrates clearly that maximum number of missing values are found for **Q1** with 75.94 % of the response data of Q1 and therefore it has been chosen to be predicted by the prediction model. Figure 4 also illustrates that the maximum number of missing responses were encountered for Q1 from HR Category Questions in every institute. Hence this Q1 has been chosen for prediction.

5.2 Predictions

After concluding that Q1 has the maximum number of missing responses among all the other questions of HR category, so the model has been developed in RapidMiner for its prediction for better vulnerability analysis.According to the system administrators of all the institutes, the cyber security training and information security awareness training are provided in their institutes and the security of personal information data of students and faculty members is also provided under HR Security Area. The prediction model developed using Naïve Bayes classification technique predicted 135 missing values for Q1 refer Table 1 based on questions abbreviated Q2-Q7 refer Table 1. Few instances of the prediction are illustrated in Table 5 which consists of

1. 30 Records as Negative Responses,22 Records as Can't Say Responses,83 Records as Positive Responses.
2. ID represented by An / Fn /Sn given in Table 5,6,7 where 'A' stands for System Administrator, 'S' stands for students, 'F' stands for faculty, 'n' stands for sequence number is used for representing the type of user giving response whether system administrator, faculty or student.
3. P(Q1) represents Predicted Responses of Q1, CY/CN/CC represents the confidence value of Yes, No, Can't Say Responses.

5.2.1 Observations from prediction model

1. The responses are "N" only when the responses of all Q2, Q3, Q4, Q5, Q6, Q7 are "N" as shown in Table 6a. It concludes that if no information security training awareness is provided to staff and students and no cyber training is provided to students and faculty members then students and faculties are sure that no security of personal information is provided.
2. Table 6b shows if any one of the training is provided then students and faculties responses are "Y" for Q1 which means that if any one training has been provided than users are aware that their personal information is also secured. It shows that students and faculties are well aware of trainings conducted and security provided in their institutes.
3. Table 6c shows the responses predicted as 'C' based on responses of Q2 to Q7 given in Table 1. Can't Say responses among students and faculty indicate the gap existing between the higher authorities and the lower level about the cyber security which leads to vulnerability of institutes.

5.2.2 Validation of the Model

Validation is performed with 65% training data and 35% testing data. It is found that the model build showed accuracy of 70.83% when the training data include all the Yes, No, Can't Say responses for question Q2, Q3, Q4, Q5, Q6, Q7 as shown below in Table 7a. The association rule that has been deduced after the prediction that if the information security training is provided then the response of question whether personal information of students or teachers is secured is yes has the support and confidence as shown in Table 7b.

5.3 Results of Vulnerability Analysis

After prediction vulnerability analysis has been done. On the basis of Theory 1 and Theory 2 mentioned in Table 2 the vulnerability percentage has been evaluated for various cases.

1. After the predictions the gap among the responses of students, faculties and system administrators has been found again for vulnerability analysis of HR category to validate if predictions have made the analysis better or not. In Figure 5 it is clearly illustrated that after the predictions the missing values are handled however there is significant number of responses giving **Can't Say** Response present for **Q1, Q2, Q3** as in Table 1**.** This indicates the gap existing in the knowledge of students and faculties about the trainings taking place in their institutes which could be due to lack of awareness of trainings, lack of proper flow of information from higher levels to lower levels about the cybersecurity trainings.
2. Figure 6. Illustrates the vulnerability percentage calculated according to Theory 2 given in Table 2 for the four categories of the Questionnaire. It has been found that out of all categories Risk Management Category is the most vulnerable out of these as shown below in Figure 12.
3. On drilling further Figure 7 shows vulnerability percentage of the questions under Risk Management category. The questions in this category are as follows

   Q1 Does your institution have a risk management program?

   Q2 Is your data backup process frequency consistent?

   Q3. Does your institution have any methods to protect and track status of media that has been removed from University website?

   Q4. Does your institution have security reviews completed at planned intervals?

   Q1 is the most vulnerable question out of all indicating that students and faculties have less awareness about risk management program.They might not be aware of the tools used for Risk Management,or the activities like reviews,backup process,recovery process as admins are the main users who deal with and implement Risk Management Programs for institute data.However it is necceassary that students and faculties should also be aware of such backup and recovery tools so as to save their work in the labs in case of system failures.
4. On further drilling down to detail level the most vulnerable institute for this category is found to be **I1** as shown in Figure 8**.** Thus, they should have more trainings regarding risk management programs. System Admins should give proper knowledge transition and take play back sessions about the various tools and the process followed for Risk Management to the students and faculties as well.
5. The above analysis was based on Theory 2 given in Table 2 but now on the basis of Theory 1 the questions with Maximum number of Can't Say responses has been identified for both admin and students and faculties respectively. The aim is to update such questions or to remove from our questionnaire. In figure 9a maximum admins responded **Can't Say** for questions from HR category and Risk Management category also. In figure 9b maximum students and faculties responded **Can't Say** for questions from Acquisition and System Maintenance Category. Thus, they will be updated or removed later.

   Below are the tables 5 to 7(a-b) and figure 3 to 9(a-b) used and mentioned in Experimental Result obtained above.

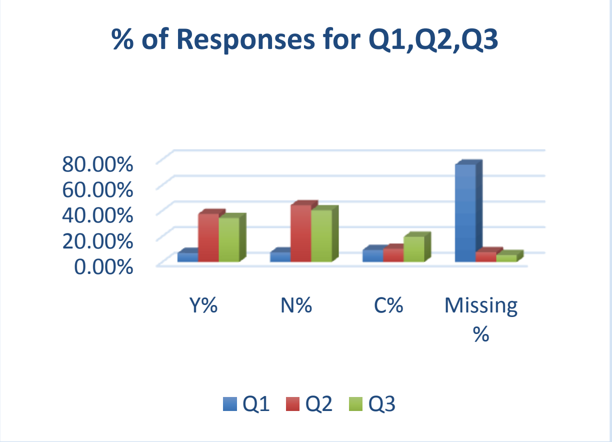

Figure 3. Responses for Questions given in Table 1 (Limited number of Questions are taken for illustration)

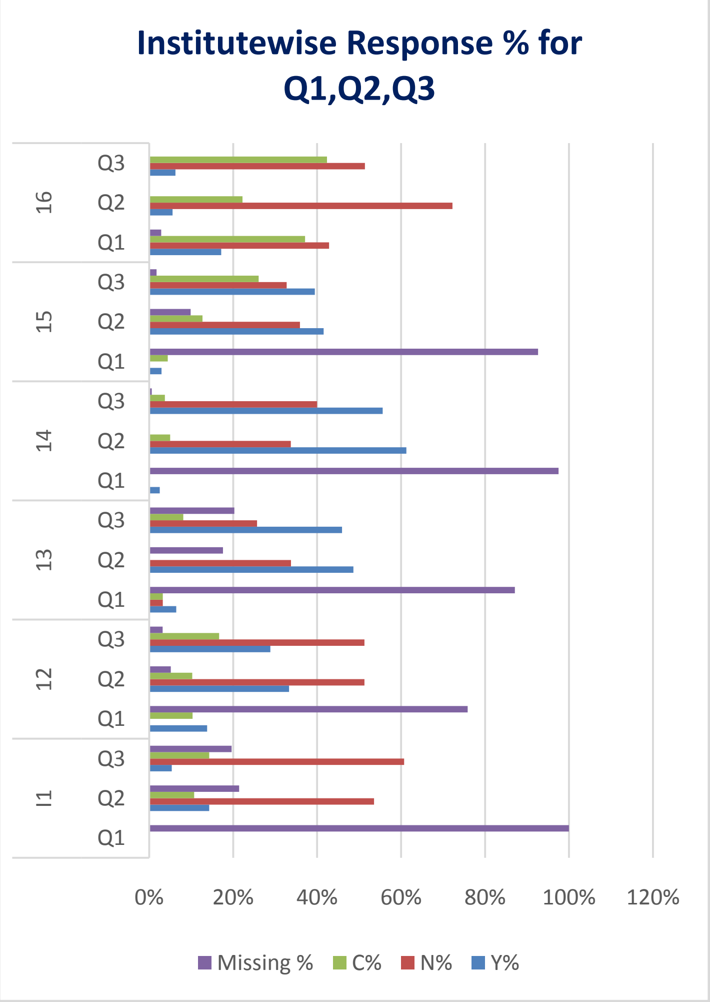

Figure 4. Institute wise Responses for Questions given in Table 1 (Instances of 6 institutes and 3 questions have been taken for illustration)

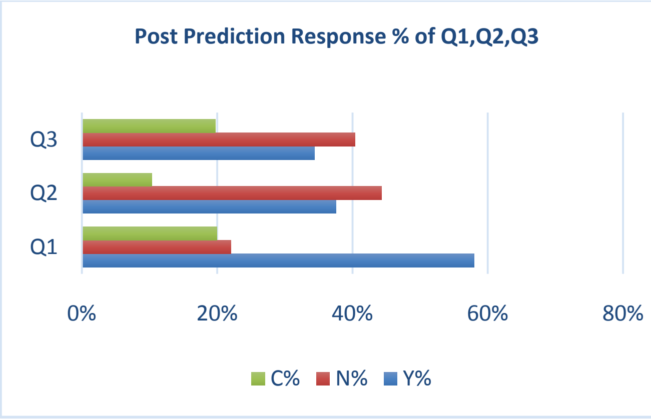

Figure 5. Responses for Questions given in Table 1 after predictions for Q1 (As an illustration)

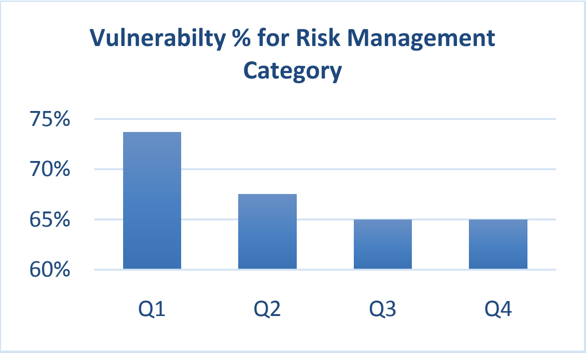

Figure 7. Vulnerability of Questions in Risk Management Category

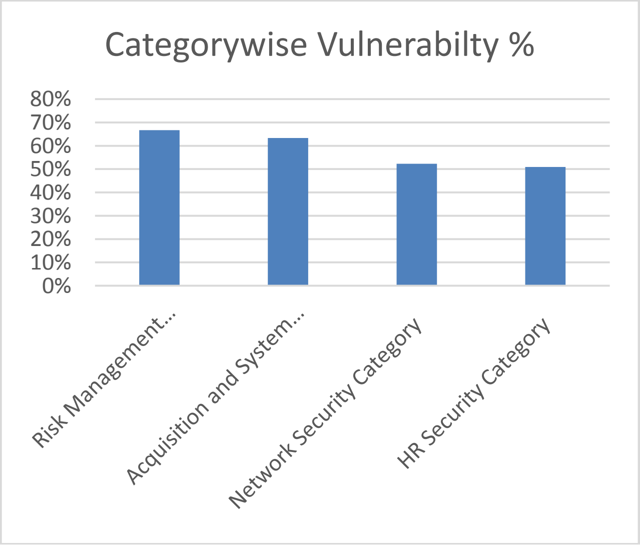

Figure 6. Vulnerability issues in proposed category of Cyber Security

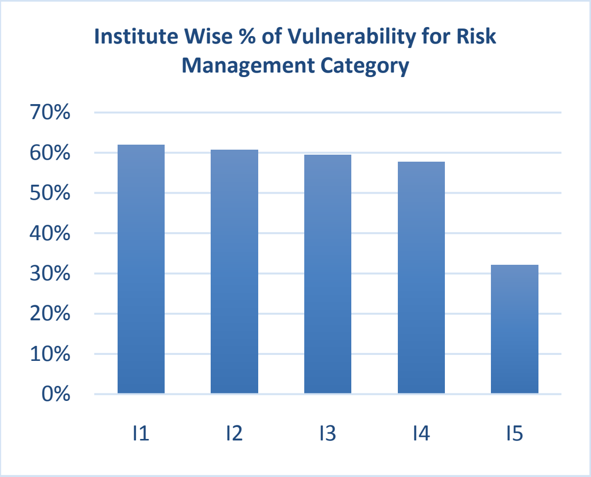

Figure 8. Institute Wise Vulnerability for Risk Management Category

Top C% of System Administrator	
Questionnaire	C%
Does your institution have data security for Research data?	42.86%
Does your institution have data security for Any other data?	40.00%
Does your institution have data security for Project data?	28.57%
Does your institution have data security for Faculty members records?	28.57%
Does your institution have data security for Student results?	28.57%
Does your institution have data security for Student records?	28.57%
Does your institution have data security for Personal Information?	28.57%
Does your institution provide e-mail ids through university portal to research scholars?	20.00%
Does your institution have any methods to protect and track status of media that has been removed from University website?	16.67%
Does your institution have any Information Security policy?	16.67%

(a)

Top C% of Students and Faculties	
Questionaire	C%
Does your institution perform Application layer vulnerability testing for critical information systems?	40.08%
Does your institution perform Network layer vulnerability testing for critical information systems?	37.55%
Does your institution perform Penetration layer vulnerability testing for critical information systems?	40.08%
Does your institution address the Code Injection application layer security vulnerabilities?	38.40%
Does your institution address the Cross-site scripting (XSS) application layer security vulnerabilities?	37.13%
Does your institution address the Cross site Request Forgery (CSRF) application layer security vulnerabilities?	38.40%
Does your institution address the Packet sniffing Network layer vulnerabilities?	36.29%
Does your institution address the DOS attacks Network layer vulnerabilities?	37.13%
Does your institution address the ICMP attacks Network layer vulnerabilities?	35.86%
Does your institution address the DDOS Network layer vulnerabilities?	35.86%

(b)

Figure 9(a-b). Top 10 Questions with maximum C% (Can't Say) response given by System Administrator (a) and Students &Faculty (b)

Table 5. Instances of Predictions for missing responses of Q1 based on Questions given in Table 1 abbreviated as Q2 TO Q7

ID	P(Q1)	CY	CN	CN	Q2	Q3	Q4	Q5	Q6	Q7
F1	C	0.003	0.071	0.926	N	N	N	C	C	C
F2	N	0.031	0.938	0.031	N	N	N	N	N	N
F3	Y	1.000	0.000	0.000	Y	Y	Y	Y	Y	Y
S1	C	0.001	0.028	0.971	C	N	N	C	C	C
S2	C	0.003	0.071	0.926	N	N	N	C	C	C
S3	N	0.031	0.938	0.031	N	N	N	N	N	N
S4	N	0.031	0.938	0.031	N	N	N	N	N	N
S5	Y	0.970	0.000	0.030	Y	Y	N	N	N	C

Table 6(a-c). Instances of Predicted Response based on Table 2 as "No/N" (a), "Yes/Y" (b) and " Can't Say/C" (c)

(a)

ID	P(Q1)	CY	CN	CN	Q2	Q3	Q4	Q5	Q6	Q7
F2	N	0.031	0.938	0.031	N	N	N	N	N	N
S3	N	0.031	0.938	0.031	N	N	N	N	N	N
S4	N	0.031	0.938	0.031	N	N	N	N	N	N

(b)

ID	P(Q1)	CY	CN	CN	Q2	Q3	Q4	Q5	Q6	Q7
F3	Y	1.000	0.000	0.000	Y	Y	Y	Y	Y	Y
S5	Y	0.970	0.000	0.030	Y	Y	N	N	N	C
S6	Y	1.000	0.000	0.000	Y	Y	Y	Y	C	N

(c)

ID	P(Q1)	CY	CN	CN	Q2	Q3	Q4	Q5	Q6	Q7
F1	C	0.003	0.071	0.926	N	N	N	C	C	C
S1	C	0.001	0.028	0.971	C	N	N	C	C	C
S2	C	0.003	0.071	0.926	N	N	N	C	C	C

Table 7 (a-b). Accuracy Model (Confusion Matrix) (a) and Support and Confidence (b)

(a)

Accuracy 70.83%	True Y	True N	True C	Class precision
pred. Y	7	0	1	87.50%
pred. N	2	4	0	66.67%
pred. C	3	1	6	60.00%
Class recall	58.33%	80.00%	85.71%	

(b)

Support	**Confidence**
0.34	0.80

6. DISCUSSION

On the analysis of above results the missing responses which were found due to lack of awareness among the students and faculty about the security policies in their institutes are predicted after the survey. These predicted responses were based on the association whether the information security awareness and cyber training has been provided or not. The results of the prediction model developed to predict whether the personal information of the staff and students is secured or not found that the students and faculties who have undergone either cybersecurity training or information security awareness training were aware of the fact that their personal information is secured and how while students who have not undergone both these trainings were not aware about the security of the personal information in their institutes. Thus, both these trainings are necessary for the students and faculties to be aware of things related to cybersecurity in their institutes. While analyzing vulnerability after predictions the results also found huge number of Can’t Say Responses of about 20% indicating the gap in the knowledge of students and faculties which could be due to lack of awareness of trainings, lack of proper flow of information from higher levels of System Administrators to lower levels of students and faculties about the cybersecurity trainings. Improper communication of the trainings schedule to students or faculties, trainings not happening at regular intervals, inefficient trainers, less importance to the trainings as compared to other academic activities can all be cited as the major reasons for this gap. In order to strengthen the cybersecurity of academic institutes it is essential to mitigate such issues by changing policies and implementing them at all levels within the institutes. Proper Communication via mails should be done about the trainings. Proper knowledge Transition and playback sessions should be scheduled for students, trainings should be made compulsory, practical knowledge should be given not just theoretical knowledge eg students should be encouraged to take backup of the data using backup recovery tools according to their policy of institutes. Thus, these measures will surely help in minimizing the gap and cyber security vulnerability of academic institutes. On the basis of the proposed research, few facts (F_n) are observed as –

F1. The students are least aware about the tools used in Cyber Security.

F2. There is knowledge gap among System Administrator, students and Faculty to aware about Cyber Security in academic institutions under different areas.

F3. There is need to evaluate the Association mining among the various responses.

F4. There is a need to find the other specific area of Cyber Security like Physical Security, Information Security etc to do more extensive Research to find vulnerabilities in Academic Institutes.

7. CONCLUSION AND FUTURE WORK

In this paper, the survey has been conducted using a questionnaire in the various academic institutes which covers various areas of cyber security. As the cybersecurity issues are increasing swiftly in the higher educational institutes thus keeping the data of users at risk and vulnerable to attackers there is a dire need to provide security guidelines in this domain. The data of the survey contained huge amount of missing values due to lack of awareness among the students, faculties about the security guidelines and policies. Hence initially, after the survey the problem of prediction has been chosen where the missing responses of question whether personal information of staff and students is secured or not are predicted for better and effective vulnerability analysis of academic institutes. These responses predicted were based on the association whether the information security awareness and cyber training has been provided or not. Our study concluded that if the training is provided in any one of these areas then the students are aware of the security of the personal information in their institutes while others lack in this awareness. Our model correctly predicted the values according to above rules deduced after the analysis. The study however also found huge number of **Can’t Say** responses indicating a gap in the responses of the admin, students and faculty and the vulnerability of the institutes in terms of cyber security. It indicates that flow of information regarding cyber securities, training from Administrators to the students and faculties is not proper and lacks efficiency. Risk Management Category is found to be the most vulnerable Category out of all discussed categories of the survey. After this detail level analysis has also been done to find which institute is the most vulnerable and which Question of the survey is found to be most vulnerable under Risk Management Category. To counter this proper information security trainings should be conducted at regular intervals. The trainers assigned should be efficient and give quality sessions. These trainings should be made mandatory like others subjects in academic curriculum and attendance should be made compulsory. Proper schedule should be communicated in well advance to students and faculties through emails, banners or pamphlets. All these guidelines will help in decreasing the gap of cybersecurity awareness among students and faculties in academic institutes.

Top 10 Questions with maximum Can’t Say Response by admins, students, faculties have also been identified. They will be updated in future work as part of designing of more effective questionnaire. Hence in future work in depth vulnerability analysis will be done for the academic institutes. This will be done by identifying new associations between the different questions based on the user response. This research also has certain limitations. This study currently has limited dataset as its data was gathered from only a few universities and colleges in Delhi Region. Thus, the research findings may be closely linked to these particular academic institutes only. However, in future, more higher education institutes will be included in the research to widen the scope of this study. This analysis will help to frame state of the art security guidelines for the academic institutes in future and make sure their compliances by the academic users.

8. REFERENCES

[1] “Why Information Security in Education?” [Online]. Available:https://nces.ed.gov/pubs98/safetech/chapter1.asp

[2] “Criminals target school data: top 10 security measuresfor educational institutions” [Online]. Available:

https://www.welivesecurity.com/2014/07/03/schools-and-cybercrime-data-security-in-education/

[3] “Cybersecurity Challenges Facing Higher Education” [Online].Available:http://www.centerdigitaled.com/higher-ed/8-Cybersecurity-Challenges-Facing-Higher-Education.html/

[4] “PM Modi asks top police officials to deal with cybercrimeon priority”[Online].Available: http://indianexpress.com/article/india/pm-narendra-modi-

asks-top-police-officials-to-deal-with-cyber-crime-on-priority-5016598/

[5] F. H. Katz (2005), ‘The effect of a university information security survey on instruction methods in information security,’ *in Proceedings of the 2nd Annual Conference on Information Security Curriculum Development*, pp. 43-48.

[6] Bhilare D.S. (2013), ‘Information security preparedness of Indian Academic Campuses with respect to global standards: a Gap Analysis’, *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, Vol 2, No.11, pp. 57-67.

[7] Burney, S.M. A. (2010), ‘Network Usage Policies for Academic Institutions’, *International Journal of Computer Applications*, Vol. 7, No. 14, pp. 6-11.

[8] Helwig M. H. (2005), ‘Global Information Assurance Certification Paper’, *SANS*.

[9] Perkins, J. (2016), ‘Access Control Policy’, ‘London School of Economics & Political Science IT Services’.

[10] [Castell, Michelle. "Mitigating online account takeovers: The case for education." *Retrieved August* 27 (2013): 2015.

[11] Arafat, Jahidul, et al. "Emergence of Robust Information Security Management Structure around the world wide Higher Education Institutions: a Multifaceted Security Solution’, *International Journal of Computer Science Issues (IJCSI)*, Vol 9, No. 2, pp. 206-214.

[12] Ahlan, Abdul Rahman, Muharman Lubis, and Arif Ridho Lubis. "Information Security Awareness at the Knowledge-Based Institution: Its Antecedents and Measures." *Procedia Computer Science* 72 (2015): 361-373.

[13] Delavari, Naeimeh, Somnuk Phon-Amnuaisuk, and M. Reza Beikzadeh. "Data mining application in higher learning institutions." *Informatics in Education* 7.1 (2008): 31-54.

[14] Mensch, Scott, and LeAnn Wilkie. "Information security activities of college students: An exploratory study." *Journal of Management Information and Decision Sciences* 14.2 (2011): 91.

[15] Kam, Hwee-Joo, et al. "Information Security Policy Compliance in Higher Education: A Neo-Institutional Perspective." *PACIS*. 2013.

[16] Saar-Tsechansky, Maytal, and Foster Provost. "Handling missing values when applying classification models." *Journal of machine learning research* 8.Jul (2007): 1623-1657.

[17] Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification." *arXiv preprint arXiv:1201.3418* (2012).

[18] Goebel, Michael, and Le Gruenwald. "A survey of data mining and knowledge discovery software tools." *ACM SIGKDD explorations newsletter* 1.1 (1999): 20-33.

[19] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *International Journal of Computer Science and Applications* 6.2 (2013): 256-261.

[20] Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems* 14.1 (2008): 1-37.

[21] Pandey, U. K. and Pal, S., “Data Mining: A prediction of performer or underperformer using classification”, *(IJCSIT) International Journal of Computer Science and Information Technology*, Vol. 2(2), 2011, 686-690, ISSN:0975-9646.

[22] Pandey, U. K. and Pal, S., “A Data Mining View on Class Room Teaching Language”, *(IJCSI) International Journal of Computer Science Issue*, Vol. 8, Issue 2, March -2011, 277-282, ISSN:1694-0814.