International Journal of Engineering & Technology, 7 (4) (2018) 2867-2873

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.9564
Research paper

Analysis of supervised and unsupervised technique for authentication dataset

Rahul K. Dubey [1] *, P. K. Nizar Banu[1]

[1] *Department of computer science, CHRIST (Deemed to be University), Bengaluru*
**Corresponding author E-mail: rahul.dubey@mca.christuniversity.in*

Abstract

Traditional methods of data storage vary from the present. These days data has become more unstructured and requires to be read contextually. Data Science provides a platform for the community to perform artificial intelligence and deep learning methodologies on large volumes of structured and unstructured data. In the era of artificial intelligence, AI is showing it's true potential by addressing social causes and automation in various industries such as automobile, medicine and smart buildings, healthcare, retail, banking, and finance service are some of the deliverables. From a variety of sources and flooding data, AI and machine learning are finding real-world adoption and applications. The nature of the data models is trial and error and is prone to change with their discoveries for the specific problem and this is the case with the different algorithms used. In this paper, we apply machine learning algorithms such as unsupervised learning k-means, bat k-means and supervised learning decision tree, k-NN, support vector machine, regression, discriminant analysis, ensemble classification for data set taken from UCI repository, phishing website, website phishing, Z- Alizadeh Sani and authentication datasets. Authentication dataset is generated for testing Single Sign-on which learns from data by training to make predictions.

Keywords: *Supervised Learning; Unsupervised Learning; Classification; Clustering; Authentication.*

1. Introduction

Authentication [9] is the mechanism used to verify the credentials of the person, to ascertain that the person is one he/she claims to be. Also, authentication can further ensure the origin and integrity of data in electronic form, for example using a digital certificate that attests to the credibility of a website. Authentication aims to cut back the likelihood of an individual attempting to access an unauthorized resource by deliberately falsifying his identity using a person's credentials. The two commonly used authentication methods are Direct registration and Single Sign-on(SSO).

Direct authentication [6] is a primary method of authentication used mostly for authentication of a user. In this method when a user visits any site, he/she has to fill the form to register and verify either through email id, phone number sometimes both. It is difficult to remember different passwords for a different site; using the same password for the entire website is also not recommended.

Single sign-on [6] has central service, which allows the user to access multiple services during a session after being authenticated by one service provider, which manages the single sign-on between multiple clients regardless of platform, technology the user is using.

Artificial Intelligence [14] is a programming computer or machine to make rational decisions as we expect the machine to behave in such a way that it is not just constrained to human guidance. Rather than a simple model, it is trained to learn from its experience from data.

Data mining and machine learning [10] are two terms we use often. Data mining is the process of finding specific information whereas machine learning deals with achieving a given task.

Machine learning is the process of training machine with an algorithm to handle large data efficiently by predictive analysis. It can be mainly classified into three; supervised learning, unsupervised learning and reinforcement learning [10].

Supervised learning [10] is the most often used machine learning algorithm that takes input data along with labels with which model is trained to generate predictions. It is further classified into regression and classification. In regression, we try to generate a curve fitting line and make subsequent predictions whereas in classification we try to segment data into classes. Regression problem is constrained to many outputs to a fixed set whereas, in classification, problems are distinct and do not constrain the number of outputs to a fixed set.

In unsupervised learning [10] data is given to the algorithm without a class label and based on the algorithm it finds structures and patterns in the data. Some of the application of unsupervised learning is the pharmaceutical industry to find which disease is plausible to happen besides diabetes, in the retail industry to predict what is the association of product customer drawn to buy often.

Statistics [18] is science of number, branch of mathematical technique, which uses data from population or sample taken from population to perform an examination and concluded. Some of the statistical methods are standard deviation, variance, regression, etc.

Reinforcement learning [14] tries to train itself over a period on the situation it is exposed to and uses its enriched knowledge to answer problems. It is trained to solve domain specific problem with an intention to maximize performance or efficiency. Reinforcement learning learns from its experience and by interacting with the environment. This process tries to reduce the involvement of human skills, which results in saving plenty of time.

Machine learning involves five steps, which are given below as in [10].

Copyright © 2018Rahul K. Dubey, P. K. Nizar Banu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Data collection: Be it of any source format, legacy data forms the basis of the prospective learning and can help suits the learning prospects of the machine.
Preprocessing data: Quality of the data can be achieved by removal of outliers and missing values, exploratory analysis can help in nutritional content of the data.
Modeling: Here is where we choose appropriate algorithm and presentation of data as a model. Further, we split the dataset into two fragments – training set and test set. The training data is employed for making the model and the test data is employed for evaluating the model.
Model evaluation: To achieve a better result and an accurate model we have to see its performance on data, which was hidden while building the model. To test the accuracy, the test dataset is employed. This decides the accuracy of the algorithm chosen depending on the result.
Improving the performance: More time has to be spent in data collection and preparation as this step might require a change to a different model or adding more variable to increase the efficiency.
Deep learning [14] is based on Artificial Neural Network (ANN) that is a human brain model, which helps to model irrational functions. To model multiple results simultaneously ANN is extremely flexible also requires a huge amount of data.
In this paper for analyzing various classification and clustering technique four data sets are taken into consideration namely authentication, phishing website, website phishing and Z- Alizadeh Sani. Classification algorithms like k-Nearest Neighbor, decision tree, regression, discriminant analysis, SVM and unsupervised method like k- means clustering is applied. Three datasets are in the field of computer security and one medical dataset is taken to compare and analyze the performance of classification and clustering algorithms.
K-Nearest Neighbor [2] is used for comparing with distance weighted k-Nearest Neighbor on unknown classification with large training sample. Results showed that values of small values of k on small sample size perform better.
Safavian & Landgrebe in [16], presents potentials, approaches, accuracy, efficiency, computational, feature selection and applications of decision tree classifier rules, search strategies and it is compared with decision tree classifier (DTC) with the neural network.
Naive Bayes in [7] is used for retrieval of information, and machine learning classification gives prominent results on textual data. It also presents variations of Naive Bayes attempting to address the limitation of Binary Independence Model (BIM).
In [15], logistic regression, discriminant analysis and Naive Bayes are compared on fifteen different datasets from UCI repository. In the experiment, it was found logistic regression had a low asymptotic error but takes time to perform better whereas Naive Bayes performs better initially as the training examples increase it results in a higher asymptotic error.
In [4], Associative Classifiers such as CBA, MCAR, MMAC, PART, and C4.5 are applied for phishing dataset which results in high performance by finding a correlation between feature and produces rules for phishing website. Experimental results also show that for same data Multi-label Classifier based Associative Classification (MCAC) outperformed Multi-class Multi-label associative classification (MMAC).
Alizadehsani et al., in [5], aims to predict least set of features required to predict if the website is phishing or legitimate site by applying two feature selection technique namely Symmetrical Uncertainty (SU) and Information Gain (IG) and two classification algorithm PART and IREP to train selected features and it is proved that IREP algorithm performed better than PART for training set with all and important feature set.
In [1], features are extracted automatically with the help of rule generation. Seventeen important features out of thirty are extracted which classifies between legitimate sites, phishing sites and suspicious sites. The most frequently appeared feature was "Request URL" and the least frequently occurred feature was "Disabling Right Click".
In [13], seventeen unique features that distinguish legitimate websites from phishing websites ones are extracted automatically from websites to identify if the website is legit or phishing using rule-based data mining technique. Among RIPPER, PRISM and CBA C4.5 algorithm outperform with respect to accuracy. Result also shows CBA has lowest error rate with 4.75%.
In [11], a model that predicts phishing attacks by the artificial neural network is presented. The features which determine the websites type keeps changing so the model is built on C++ technology. The dataset is comprised of 800 phishing websites and 600 legitimate websites for training. Continuous training will help to improve results by modifying the learning rate before attaching new neuron to the hidden layer.
In [12], Presence of cardiovascular diseases wiz. coronary artery disease (CAD) was predicted by applying several algorithms, and feature selection techniques which resulted in high accuracy value on Z- Alizadeh Sani dataset. With thr help of this CAD can be identified with high probability in low cost. SMO algorithm performed best94.08% in comparison with Bagging, Naive Bayes and Neural network.
In [13], combined information gain and average information gain are used for feature selection. Both methods had twenty-four features which gave higher accuracy also tweaked SVM algorithm kernels with polynomial, sigmoid, linear and RBF for prediction shows high confidence level between features.

2. Methodology

The methodology followed in this paper is shown in Fig. 1 First, the data sets undergo normalization using min-max normalization technique followed by supervised and unsupervised learning. The data set applied with 10 folds cross-validation.

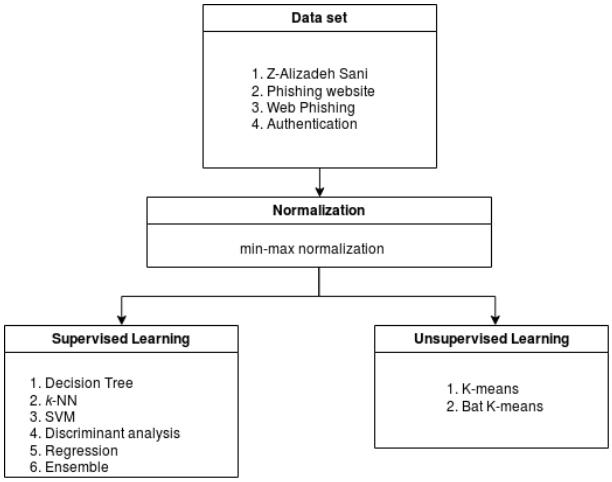


Fig. 1: Framework Followed in This Paper.

Classification[10] is the technique used for identifying a function that explains and discriminates instance classes, for the objective of being able to intelligently employ the function to predict the class of instances whose class label is undisclosed. The derived function is built on the analysis of training data set whose class label is known.

2.1. k- Nearest neighbor

K-NN [10] is one of the simplest classification algorithm. It is often used in classification which saves all instances and classifies new instance by voting for the nearest neighbor measured by distance function whereas in regression problem nearest neighbor is measured by average. It is computationally expensive as it requires iterative process and if outliers and noises are not removed in preprocessing step, it may result in bias. The pseudo code for k-NN is shown in Fig. 2 [19].

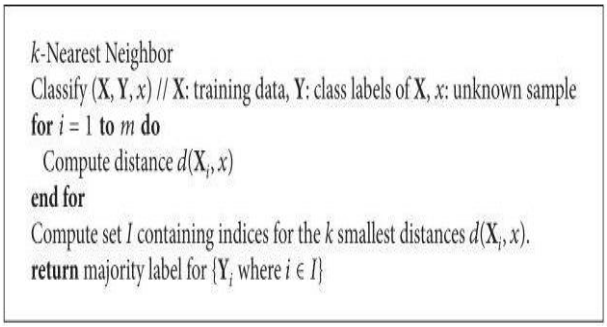


Fig. 2: Pseudo code for k-NN.

2.2. Decision tree

Decision tree [10], is a type of tree that categorize instances by arranging them based on attribute values. A tree consists of node and branch. Where every node in a decision tree represents features in an instance to be categorized, and tree presumes that the node of every branch represents a value. The pseudo code for Decision Tree is shown in Fig. 3 [17]

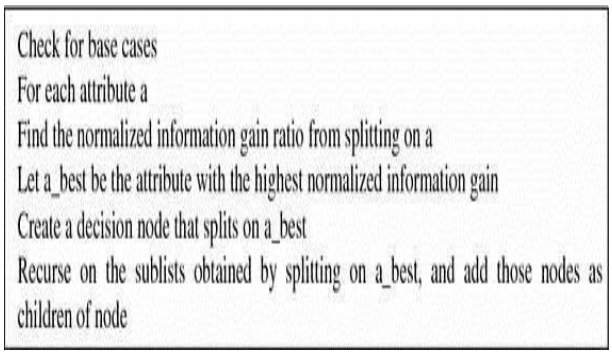


Fig. 3: Pseudocode for Decision Tree.

2.3. Naive bayes

Naive Bayes classifier [10] is a probabilistic supervised learning that categorizes based on applying Bayes theorem with strong independence hypothesis. The probability of data instance X having the class label C_j is:

$$\frac{P(C_j|X) = P(X|C_j) * P(C_j)\ (1)}{P(X)} \quad (1)$$

The class label C_j with highest conditional probability value decides the category of the data instance.

2.4. Support vector machine

SVM [18] algorithm can be applied to both regression and classification problem. All data instances are plotted in n-dimensional space where n is the number of class and each class consists certain value of coordinate. It uses a subset of training point as decision function. SVM works well with a clear margin of separation and in high dimensional spaces. It does not work well if the dataset has noise and response class is overlapping on the hyper plane.

2.5. Regression

Modeling technique which tries to find the relation between dependent and independent variable. Mapping relation between target variable and prediction variable, applications of regression analysis are forecasting, time series analysis etc. This is achieved by drawing a curve or line and minimizing the distance between the curve or line and the data points. Regression [18] are innumerable types but driven mainly based on three properties. Type of dependent variable(Y), number of independent variables(X), the shape of the line. Most widely used regressions are

Linear regression [18], the shape of the curve or line is linear, the dependent variable is continuous in nature and independent variable can be either continuous or discrete in form.The relationship between Y and X using regression line. It is of the form Y=m*X+c where m = intercept and slope of line and c is error term.

Logistic regression [18] helps to determine an event of occurrence or nonoccurrence. This technique is applied if the dependent variable (Y) is of the binary form. The value of dependent variable can range from [0] to [1] wiz. Odds are equal to probability of occurrence divided by probability of nonoccurrence can be represented as:

$$odds = \frac{p}{1-p} \quad (2)$$

$$\text{logit(p)} = \ln(\text{odds}) = \ln\left(\frac{\text{p}}{1-\text{p}}\right) \quad (3)$$

$$logit(p) = ln\left(\frac{p}{1-p}\right) = b0 + b1X1 + b2X2 + b3X3 \ldots + bkXk \quad (4)$$

2.6. Ensemble classifier

Ensemble classifier [18] is used in the creation of a robust system with help of base learning mechanism created by combining two or more algorithm of same or different types. In other words from input set, different mapping functions are applied to achieve the desired outcome. The classifier combined for ensemble classifier is boosting, random forest, bagging, random subspace, and ECOC. There are countless types of ensemble models commonly used ones are stacking, boosting and bagging.

2.7. Dimensionality reduction

Dimensionality reduction is a mechanism of transforming a vast number of features in a data set into a dataset with a lesser dimension which does not lose any information in a succinct manner. It reduces storage space by compressing the data and faster execution of the algorithm. There are many methods to reduce the features of training data namely factor analysis, low variance, higher correlation, backward or forward feature selection and others, the commonly used methods are listed below:

Principal Component Analysis [10], tries to map dataset variables to a new set of variables which represent principal component resembles like a linear collection of the original variable. The first principle component has most of the combination of original data. The second and the following principles intersect in an orthogonal fashion. PCA is not suggested if it loses its meaning, if the variables are not standardized or if the result requires additional explanation.

Discriminant analysis [10], is a dimensionality reduction technique by authors, but the discriminant analysis also works as a classifier. Two commonly used discriminant functions are Linear Discriminant Analysis and Quadratic Discriminant Analysis.

Clustering is the process of grouping the instances into different groups. This way the data points in the same cluster are more alike to other instances in the same cluster than those in other clusters. In other words, the aim is to separate cluster with similar properties and assign them into clusters.

2.8. k- means

It is Centroid-Based Technique. k- means [10] takes two arguments in the algorithm. First is the input data with n objects and second is the number of clusters, it is an iterative process that aims to find local maxima by assigning each point closest to cluster centroid and recomputing cluster centroid and the algorithm ter-

minates if no change is found in movement of object. The pseudo code for k- means [20] is shown in Fig. 4.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;

Fig. 4: Pseudocode for K- Means.

2.9. Bat k- means

Bat k-means is a problem independent to technique to the problem it uses echolocation to find the distance between object and cluster centre by randomly assigning objects to k clusters. The pseudo code for bat k- means [21] is shown in Fig. 5. presented by (Tang et al. 2012) is used for our experiments.

Algorithm : **Bat K-Means**
Input : Jaundice dataset J_{mXn}, K – Number of clusters $(K \leq m)$, frequency factor Q *and* loudness A
Output : Clusters $C_1, C_2, \ldots, C_K$ and Centroid of clusters $Z_1, Z_2, \ldots, Z_K$

Procedure:

1. Randomly assign K clusters for each of the N bats
2. For each bat, select K objects from S data objects as initial centroids, by taking the mean values of the attribute of the objects within their given clusters
3. Calculate the fitness of the centroid in each bat, and find the best solution that is represented by the total fitness values of centroid in a bat
4. Generate a new solution by adjusting the frequency, updating the velocity and creating new centroid values
5. If random[0..1]>pulse rate R
   5.1 For each bat, select a solution among a set of best solutions from the other bats, and generate a new local solution around the selected best solution, else goto step 9
6. If random[0..1]<A_i and $f(x_i)<f(x_j)$ else goto step 9
7. Accept new solutions, increase R_i and reduce A_i
8. Reassign the clusters
9. Output the best cluster configuration that is represented by the bat that has the greatest fitness
10. Repeat steps 2 to 9 until convergence

Fig. 5: Pseudocode for Bat K- Means.

3. Experimental analysis

3.1. Dataset description

This section discusses dataset used for prediction purpose and Table. 1. shows its description.

Table 1: Dataset Description

s. no	Dataset	Instances	Features	No. of Classes	Class name
1	Z- Alizaiden Sani	303	55	2	CAD, Normal
2	Phishing website	11055	30	2	Phishing site, Legit site
3	Website phishing	1353	10	3	Phishing site, Legit site, Suspicious site
4	Authentication	100	13	2	Multiple accounts, Single Sign-on

Z- Alizadeh Sani [4], [5] dataset is taken from http://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani. The dataset does not contain any missing values. It has a total of 54 attributes and303 instances. The attributes are classified into four categories as demographic, symptom and examination, ECG, and laboratory and echo. There are two classes in which a patient is categorized as Normal or CAD. Normal, if patient diameter narrowed is lesser than 50% else the patient is of a class CAD. A sample of ten attributes is shown in Table. 2.

Table 2: Attributes In Z- Alizadeh Sani Dataset

s. no	Features
1	Age
2	Weight
3	Length
4	Sex
5	BMI
6	DM
7	HTN
8	Current Smoker
9	EX-Smoker
10	FH

Phishing Websites [12], [13], [14] dataset is taken from https://archive.ics.uci.edu/ml/datasets/Phishing+Websites. In the field of computer security, due to lack of availability of reliable training dataset researchers in that characterizes phishing websites, the creator of the data set has focused on important attributes that focus on identifying phishing websites. Dataset consists of 30 attributes with integer characteristics and 2456 instances. The class labels are legitimate, suspicious and phishy represented in numerical values as [1-0] and – [1] respectively. There were no missing values. A sample of 10 attributes is shown in Table. 3.

Table 3: Sample of Attributes in Phishing Website Dataset

s. no	Features
1	having_IP_Address
2	URL_Length
3	Shortining_Service
4	having_At_Symbol
5	double_slash_redirecting
6	Prefix_Suffix
7	having_Sub_Domain
8	SSLfinal_State
9	Domain_registeration_length
10	Favicon

Website phishing [15] [16] data set is taken from the website https://archive.ics.uci.edu/ml/datasets/Website+Phishing consisting of different sources that identifies a website is legit or not. It has become a vital issue which involves monetary aspects. The class labels are Legitimate, Suspicious and Phishy represented in numerical values as 1,0 and -1 respectively. There were no missing values. The phishing website was collected from Phishtank data archive www.phishtank.com and PHP script was plugged in the browser to collect 1353 websites containing 548 legitimate websites. There is 103 suspicious URLs and 702 phishing URLs. This dataset consists of ten predictors and one response class shown in Table. 4.

Table 4: List of Attributes in Website Phishing Dataset

s. no	Features
1	URL Anchor
2	Request URL
3	SFH
4	URL Length
5	Having
6	Prefix/Suffix
7	IP
8	Sub Domain
9	Web traffic
10	Domain age
11	Class

Authentication Dataset has a population of people studying in CHRIST (Deemed to be University), Bangalore, India below the age of 25 and are Undergraduates and Postgraduates. Data is collected between July 25, 2017, and September 7, 2017. The sample was done using purposive sampling. Since the class labels were known in advance, supervised learning techniques are used for classification. This research uses Primary data which employs the use of the structured survey in extracting the essential data to experiment. The research instrument consists of

1) A brief set of demographic questions, including questions on age and sex.
2) Questions to obtain information on the behavior of choosing authentication model.

Table. 5. Shows the questions asked to the survey participants and their corresponding response data type. A sample of hundred respondents was adopted to conduct the research. There were in total 15 features out of which 12 were taken into consideration, which had 0% missing value.

Table 5: Features and Data Type in Authentication Dataset

s. no	Questions/ Features	Answer type
1	Which of the option do you think is preferable and safe for your daily usage?	Nominal{email ID Password, SSO }
2	Do you wish to have multiple accounts or prefer to use a single account for all the sites?	Binary Nominal
3	How often do you tend to forget multiple username and passwords?	Ordinal {1- often, 5- never}
4	Choose one or more factors mentioned below that will be considered affecting your decision to choosing to use SSO while comparing it with direct authentication method	Nominal {Familiarity and ease of use, Convenience and time saving, All of the above}
5	How awful will you feel if your personal information provided to register/ sign account details were at risk?	Ordinal {1- Low, 5- high}
6	As an end user are you likely to not use (not register/not sign on) a website if it does not allow you to use single sign-on method?	Binary {Yes, No}
7	How critical/informative is your understanding about the importance of safety and risk involved in using your details to sign on to a website or subscription using direct registration method?	Nominal
8	As a business owner if you were to integrate single sign-on method instead of a direct registration to your online business, do you think you are more likely to protect and retain your customer's identity and your business integrity.	Binary {Yes, No}
9	According to you which is a most important attribute to ensure secure authentication mechanism and information security?	Nominal {Use strong passwords example: (P60&Nk@b), Use plugins that offer an extra layer of security, Keep yourself updated by with operating system and antivirus software}
10	What one or more action will you take if you receive a notification mail for a login activity from an unrecognized operating system, location and time?	Nominal {Change Password for account, Report before changing the password for the account, Delete the account, Use password valet to secure your passwords}
11	How important do you think the service provider should be serious about security and take necessary steps for its user privacy?	Ordinal {1-Least, 5- Always }
12	Likelihood of raising awareness about information security to your family and friends	Ordinal {1- Likely, 5- Unlikely}

3.2. Comparative analysis

In this section, Fig.6. – Fig. 12. Shows the accuracy percentage of different algorithm namely decision tree, SVM, regression, ensemble, KNN and discriminant analysis experiment results are discussed for data as follow dataset 1 represents Z- Alizadeh Sani, 2 represents Phishing website, 3 represents website phishing and 4 represents authentication data.

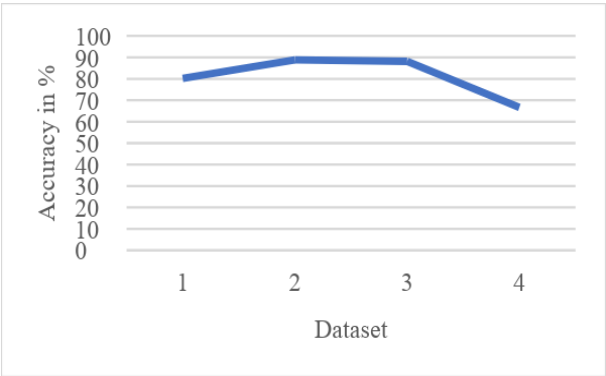


Fig. 6: Accuracy of Decision Tree Classifier.

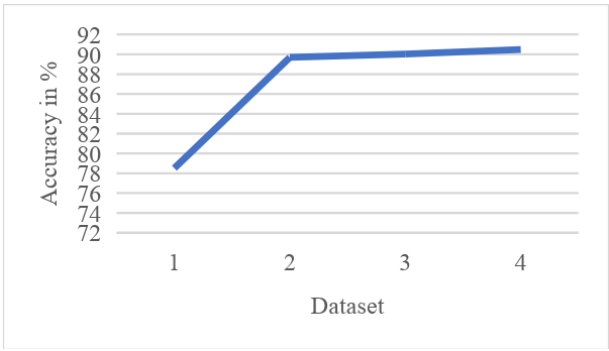


Fig. 7: Accuracy of Support Vector Machine.

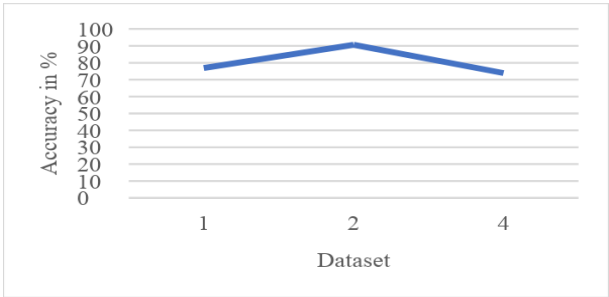


Fig. 8: Accuracy of Regression.



Fig. 9: Accuracy of Ensemble Classifier.

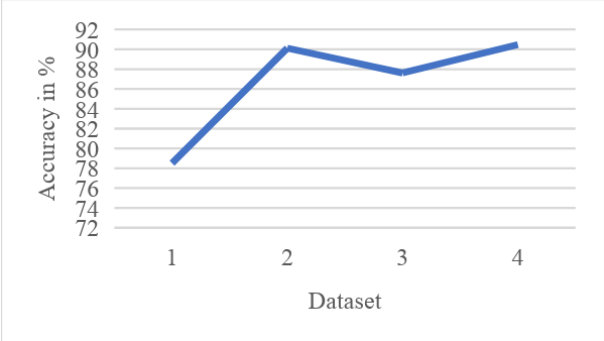


Fig. 10: Accuracy of k-NN.



Fig. 11: Accuracy of Linear Discriminant.

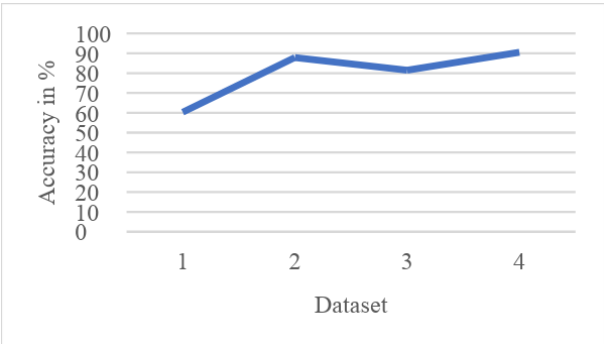


Fig. 12: Accuracy of Quadratic Discriminant.

Accuracy of various classifiers applied on authentication dataset is shown in Table. 6. Fig. 13. Shows accuracy of Decision Tree (DT), SVM (Support Vector Machine), k- Nearest Neighbor, Ensemble, Linear Discriminant (LD), and Quadratic Discriminant (QD).

Table 6: Accuracy of Algorithms on Authentication Data Set

s. no	Algorithm	Accuracy (%)
1	Decision Tree	66.6666
2	SVM	90.4762
3	Regression	73.9130
4	k-NN	90.4761
5	Ensemble	90.4761
6	Linear Discriminant	90.4761
7	Quadratic Discriminant	90.4761

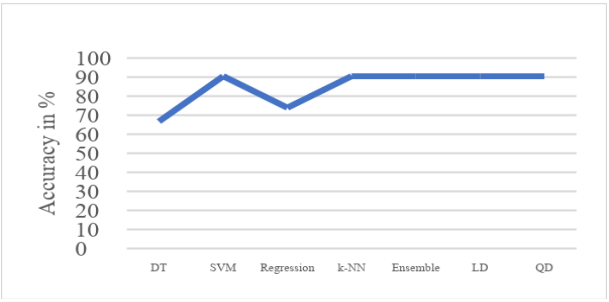


Fig. 13: Accuracy of Different Algorithm.

Table. 7. and Fig. 14. Shows the Davis Bouldin indexof k- means for all the data sets

Table 7:DB Index of Algorithms on Authentication Data Set

s. no	Dataset name	No. of cluster	DB index
1	Authentication	2	2.4722
2	Website Phishing	3	1.7708
3	Phishing website	3	1.9958
4	Z- Alizadeh Sani	2	0.6189

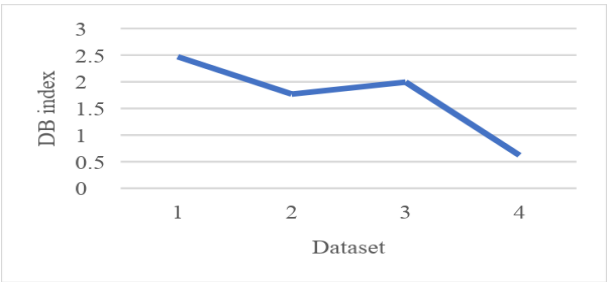


Fig. 14: Deindex of k-Means on Data Sets.

From the experiment, it was found that accuracy of decision tree and regression is not as good as other algorithm as the dataset contains 60% of instances in Single Sign-on class so decision tree using split rules resulted in 66.66% accuracy and regression could not fit the curve with 73.92% accuracy. In Fig. 8 accuracy of Regression classifier is shown only for three datasets because the variant of regression used is applicable only for two class problem since website phishing has three classes multinomial logistic regression can be applied.

The k- means clustering for authentication dataset as two classifiers, which are applied in our experiments failed to give good results compared to other classifiers. Cluster validity index namely Davis Bouldin index [8] is applied for the clusters formed using k-means clustering algorithm. It is found the error rate for authentication dataset is 2.472. k- Means is applied for all other datasets discussed in this paper. The results are shown in Table. 7. and Fig. 6. We also tried bat k- means on all the datasets represented respectively in Fig.14- Fig. 17

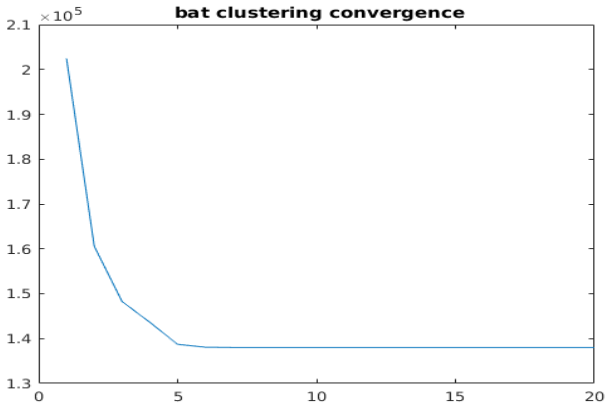


Fig. 14: Convergence of Bat K- Means for Authentication Dataset.

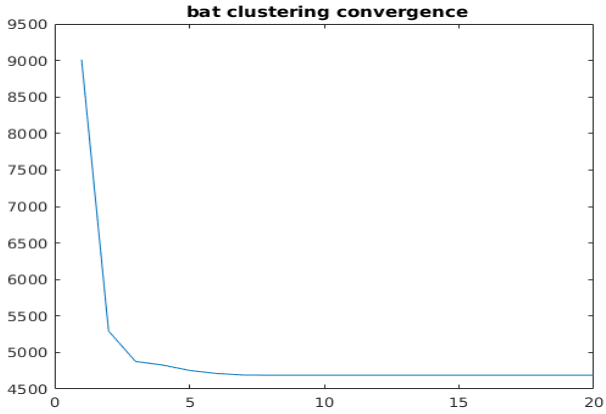


Fig. 15: Convergence of Bat K- Means for Website Phishing Dataset.

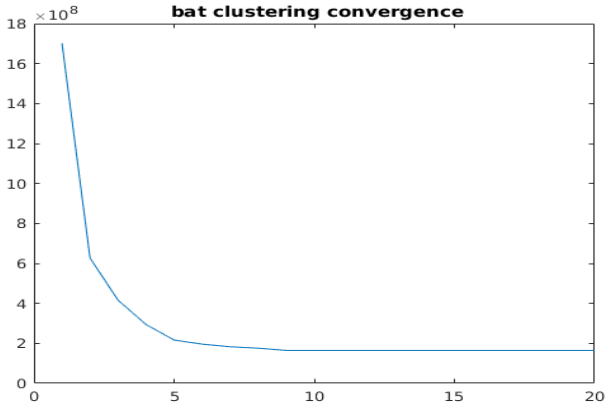


Fig. 16: Convergence of Bat K- Means for Phishing Website Dataset.

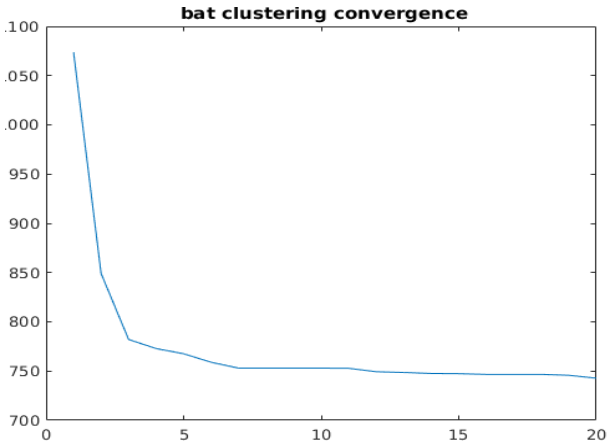


Fig. 17: Convergence of Bat K- Means for Z Alizaden Sani Dataset.

4. Conclusion

To analyze authentication dataset collected in CHRIST (Deemed to be University), a set of different classifiers are applied The accuracy of classifiers are also tested on three different datasets taken from UCI repository. Experimental results show that linear discriminant works well for Z- Alizadeh Sani dataset, Ensemble classifier works well for Phishing website and Website phishing dataset. With respect to authentication dataset k-NN, ensemble classifier, linear discriminant and quadratic discriminant gave exactly 90.47% accuracy. We tried applying k- means clustering along with classifiers for all the datasets discussed in this paper. With the experiments carried out for the authentication dataset it is understood that users prefer Single Sign-on method for registration and continual usage of services provided by various websites.

References

[1] D. (Turner), "Digital Authentication - the basics", *Cryptomathic.com*, 2016. [Online]. Available: https://www.cryptomathic.com/news-events/blog/digital-authentication-the-basics.

[2] "Authentication Patterns", *Msdn.microsoft.com*, 2015. [Online]. Available: https://msdn.microsoft.com/en-us/library/ff647374.aspx.

[3] N. Buduma, *Fundamentals of Deep Learning*, 1st ed. O'Reilly Media, Inc., 2017.

[4] J. Han and M. Kamber, *Data mining*, second ed. Amsterdam: Elsevier, Morgan Kaufmann, 2006.

[5] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, 2nd ed. Springer Science & Business Media, 2009, 2009.https://doi.org/10.1007/978-0-387-84858-7.

[6] "A Note on Distance-Weighted k-Nearest Neighbor Rules", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 4, pp. 311-313, 1978.https://doi.org/10.1109/TSMC.1978.4309958.

[7] S. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660-674, 1991.https://doi.org/10.1109/21.97458.

[8] D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval", *Machine Learning: ECML-98*, pp. 4-15, 1998.

[9] Ng, A.Y. and Jordan, M.I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

[10] Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B. and Sani, Z.A., 2013. A data mining approach for diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *111*(1), pp.52-61.https://doi.org/10.1016/j.cmpb.2013.03.004.

[11] Alizadehsani, R., Zangooei, M.H., Hosseini, M.J., Habibi, J., Khosravi, A., Roshanzamir, M., Khozeimeh, F., Sarrafzadegan, N. and Nahavandi, S., 2016. Coronary artery disease detection using computational intelligence methods. *Knowledge-Based Systems*, *109*, pp.187-197.https://doi.org/10.1016/j.knosys.2016.07.004.

[12] Abdelhamid, N., Ayesh, A. and Thabtah, F., 2014. Phishing detection based Associative Classification data mining. *Expert Systems with Applications*, *41*(13), pp.5948-5959.https://doi.org/10.1016/j.eswa.2014.03.019.

[13] Al-diabat, M., 2016. Detection and Prediction of Phishing Websites using Classification Mining Techniques. *International Journal of Computer Applications*, *147*(5).

[14] Mohammad, R.M., Thabtah, F. and McCluskey, L., 2012, December. An assessment of features related to phishing websites using an automated technique. In *Internet Technology and Secured Transactions, 2012 International Conference for* (pp. 492-497). IEEE.

[15] Mohammad, R.M., Thabtah, F. and McCluskey, L., 2014. Intelligent rule-based phishing websites classification. *IET Information Security*, *8*(3), pp.153-160.https://doi.org/10.1049/iet-ifs.2013.0202.

[16] Mohammad, R.M., Thabtah, F. and McCluskey, L., 2014. Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, *25*(2), pp.443-458.https://doi.org/10.1007/s00521-013-1490-z.

[17] Tay, B., Hyun, J.K. and Oh, S., 2014. A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images. *Computational and mathematical methods in medicine*, *2014*.

[18] Sharef, N.M., Martin, T., Kasmiran, K.A., Mustapha, A., Sulaiman, M.N. and Azmi-Murad, M.A., 2015. A comparative study of evolving fuzzy grammar and machine learning techniques for text categorization. *Soft Computing*, *19*(6), pp.1701-1714.https://doi.org/10.1007/s00500-014-1358-x.

[19] Zhou, P.Y. and Chan, K.C., 2014, May. A Model-Based Multivariate Time Series Clustering Algorithm. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 805-817). Springer, Cham.

[20] Davies, D.L. and Bouldin, D.W., 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), pp.224-227.https://doi.org/10.1109/TPAMI.1979.4766909.

[21] Banu, P., Own, H., Olariu, T. and Olariu, I. (2017). Cluster Analysis for European Neonatal Jaundice. *Soft Computing Applications*, pp.408-419.