IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 8, August 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908026 |

Performance Evaluation of Data Mining Classification Techniques

Dr. Sanjay Kumar[1], Mrs. Raksha Shukla[2]

Dept. of Computer Science, Kalinga University, Raipur (C.G.) India[1]

Ph. D. Scholar (Computer Science) Kalinga University, Raipur (C.G.) India [2]

ABSTRACT: In present era large amount of data is generated day by day. All those data are needed to be study and analyze. Data mining can be applied in different data like scientific, medical, space, aeronautics bank data analysis and so on. There are so many techniques and algorithms in data mining for this purpose such as machine learning classification, clustering, regression, ANN etc. Data mining studied data patterns and provide meaning full information.

Main aim of this paper is study of different classification techniques and evaluation of their performance. Weather data is used for this study. Classification techniques are used for getting information from those weather data.

KEYWORDS: Data Mining, Classification, C4.5, Naïve Bayes, SVM, Random Forest, KNN.

I. INTRODUCTION

Data Mining should have been more appropriately named "Knowledge mining from data" as per Han and Kamber [11]. Many other terms have a similar mining to data mining for example- knowledge mining from data, knowledge extraction, data pattern analysis, data archaeology and data dredging.KDD (Knowledge discovery in data) process.

KDD process

(1) Data cleaning (to remove noise and inconsistent data)
(2) Data Integration (multiple data sources may be combined)
(3) Data Selection (where data relevant to the analysis task are retrieved from the database)
(4) Data Transformation (where data are transformed and consolidated in appropriate forms)
(5) Data Mining (an essential process where intelligent methods are applied to extract data patterns)
(6) Pattern Evaluation (to identify the truly interesting pattern representing knowledge based on interesting pattern)
(7) Knowledge Presentation (visualization and knowledge representation techniques are used to present mined knowledge to users)

II. LITERATURE REVIEW

Data mining is most widely used method for research in present era. Most of researchers use this technique for their study and get better result. Many research work has been studied related to the topic, some of them discussed here-

Santos [1] used classification techniques for identify fake review s about the products which is helpful to analyze customers like and dislike. Educational data classification and prediction has been done [2]. They compare SVM, Naïve Bayes and Random Forest and find the Random Forest is best for their research. Mumine [3] have found that random forest and simple CART have great accuracy to detect early stage cancer. The author of [4] paper made a clinical decision support system for prediction of multiple disease. They used 25 classifiers and find some of them are suitable for prediction such as CF, LDA, GLM, RF, GP. Research paper [5] used many classification techniques, but find that gradient boost and SVM is best of all for mobile price classification. Paper [6] is about overview of different classification techniques and their merits and demerits as well. Paper [7] discussed about hybrid model of different neural network optimization algorithms such as BPNN, RNN, LM and find appropriate classification. In [8] author made IDS system using hybrid form of classification algorithms. Random forest is best for IDS comparison to others. By the [9] authors utilizes the classification algorithms NN,NB, RF, KNN and find Random Forest gives more accuracy than others for weather prediction. In [10] researchers used KNN, RF, NB and DT but find KNN is best for soil data classification.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908026 |

III. METHODOLOGY

A. *Classification*

Classification is a main domain of data mining technique which maps data into predefined groups or classes. It is supervised learning because classes are defined before examining of data.

Classification is used for different purposes like machine learning, pattern recognition, network security, medical science etc. There are many classification techniques decision tree, Naïve Bayes, KNN, SVM etc.

1) C4.5 Algorithm:

C4.5 is successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a "split information" value defined analogously with info (D) as

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

This value represents the potential information generated by splitting the training dataset, D, into v partitions, corresponding to the v outcomes of a test on attribute A. The gain ratio is defined as:

$$Gainratio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

The attribute with the maximum gain ratio is selected as the splitting attribute.

2) Naïve Bayes:

Naïve Bayes is the supervised machine learning algorithm. It can be used for both classification and regression. It is probabilistic classifier. It requires small number of training data for classification.

Bayes' Theorem:

Probability (B given A) = (Probability (A and B) / Probability (A))

3) Random Forest:

Random Forest [11] is an ensemble method. It is a decision tree classifier so that the collection of classifiers is a "forest". The individual decision trees are generated using a random selection of attributes at each node.

Random Forest is a supervised learning algorithm, describe as a combination of a tree predictors. It is used for both classification and regression. It is most accurate general purpose learning technique.

4) SVM:

Support Vector Machine is a supervised learning algorithm. It is a method for the classification of both linear and nonlinear data. It uses nonlinear mapping to transform the original training data into a higher dimension.

The training time of fastest SVM [11] can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries. They are much less prone to over fitting than other methods. SVM also provide numeric prediction as well as classification. They have been applied to a number of areas, including handwritten digit recognition, object recognition and speaker identification, as well as benchmark time- series prediction tests.

5) kNN:

The k-nearest-neighbor [11] method was first described in the early 1950s. The method is labor intensive when given large training sets. It has since been widely used in the area of pattern recognition.

Nearest-neighbor classifiers are based on learning by anology, that is by comparing a given test tuples with training tuples that are similar to it. The training touples are described by n attributes. Each touple represents a point in an n-dimensional space. In this way, all the training touples are stored in an n-dimentional pattenr space. When given an unknown touple, a k-nearest-neighbor classifier searches the pattern space for the k training touples that are closest to the unknown space. These k training touples are the k "nearest neighbors" of the unknown touple.

"Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or touples, say, $X_1=(x_{11},x_{12},.....,x_{1n})$ and $X_2=(x_{21},x_{22},.....,x_{2n})$, is

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908026 |

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

B. Weka

Waikato Environment for Knowledge Analysis (Weka), as given in Wikipedia[12] is a data mining/machine learning tool developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License, Weka is a bird found only in New Zealand. This is a collection of machine learning algorithms for data mining tasks. The algorithm can either be applied directly to a dataset or called from Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. It is also well suited for developing new machine learning schemes. Weka provides access to SQL databases using Java database Connectivity and can process the result returned by a database query.

C. Dataset

Weather data is used for this research paper. The dataset has been taken from UCI Machine learning repository. It is 1(2013) year weather data for Washington City so there is 366 instances. It consist three attributes precipitation, temperature and wind.

IV. PERFORMANCE METRICES

In this section we discuss about how accurate any classifier predicting class label of touples. Performance of the classifier include accuracy, sensitivity (or recall), specificity, precision, F_1 and F_β. Accuracy is a specific measure so it can be used to check classifier predictive ability. According to [13]-

Accuracy (CA) refers to the correct predictions rate. It is given by the division of total correct predictions by the total number of instances.

$$CA = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is also called positive predictive value and reports which of those who were predicted to be positive are actually positive. It is defined as the number of true positives divided by the number of true positives plus the false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Sensitivity calculates the true positive rate, and this is how many of the actual positives were correctly labeled. It is defined as the number of true positives divided by the number of true positives plus the number of false negatives.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity defines the true negative rate. This is the proportion of actual negatives which were correctly predicted. To obtain this metric, we divide the number of false positives by the number of true negatives plus the number of false positives.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

F1-Score is the harmonic mean of Precision and Recall. It presents a better measure of the incorrectly classified cases than the CA.

$$\text{F1} = 2.\frac{precision.sensitivity}{precision + sensitivity}$$

V. RESULT AND EVALUATION

This section discuss about result of five classification algorithms. The experiments are applied in 365 weather dataset. We can see performances of those algorithms like time taken to build up model, accuracy, precision,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908026 |

sensitivity (recall), F_1 score etc. the screenshots of result of Naïve Bayes, C4.5, Random Forest, SVM and kNN algorithms are shown bellow

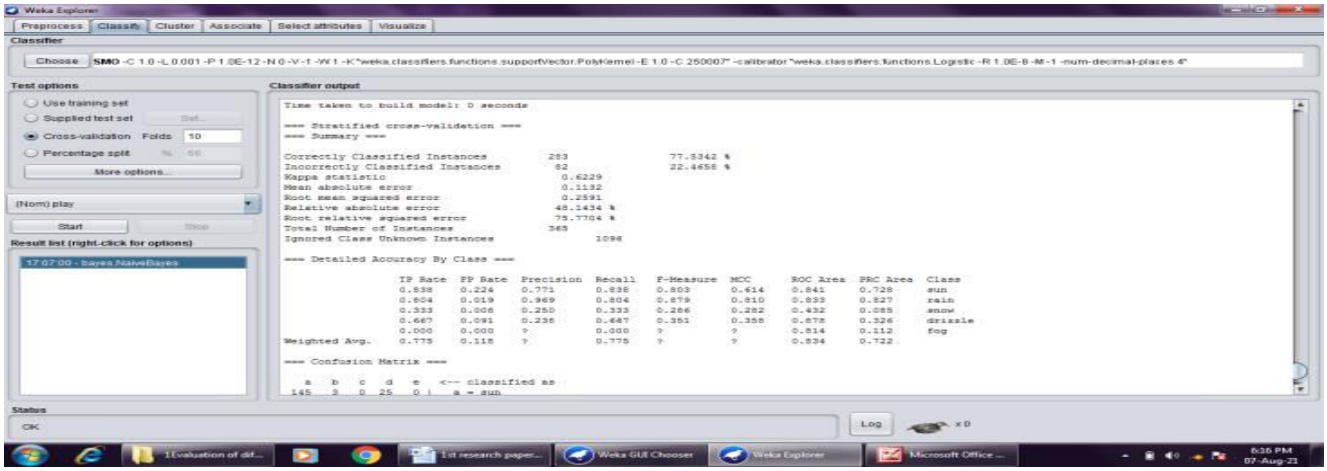

Fig.1 Result of Naïve Bayes

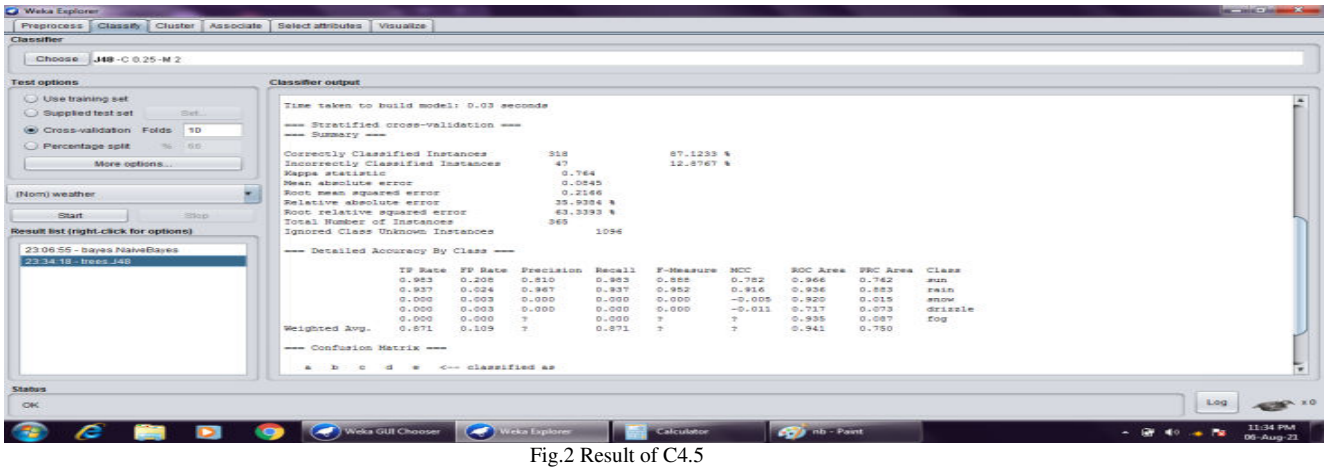

Fig.2 Result of C4.5

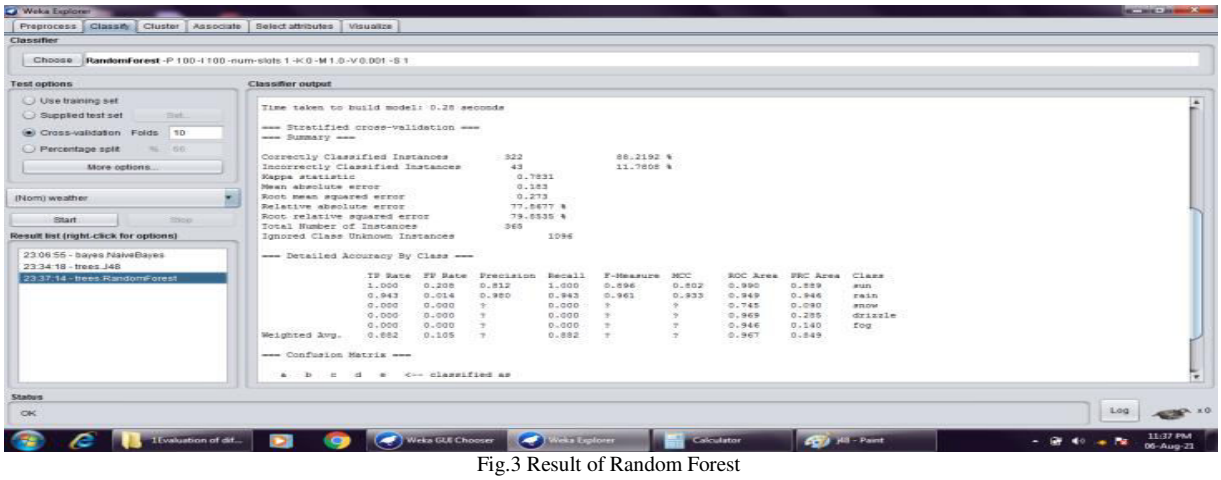

Fig.3 Result of Random Forest

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908026 |

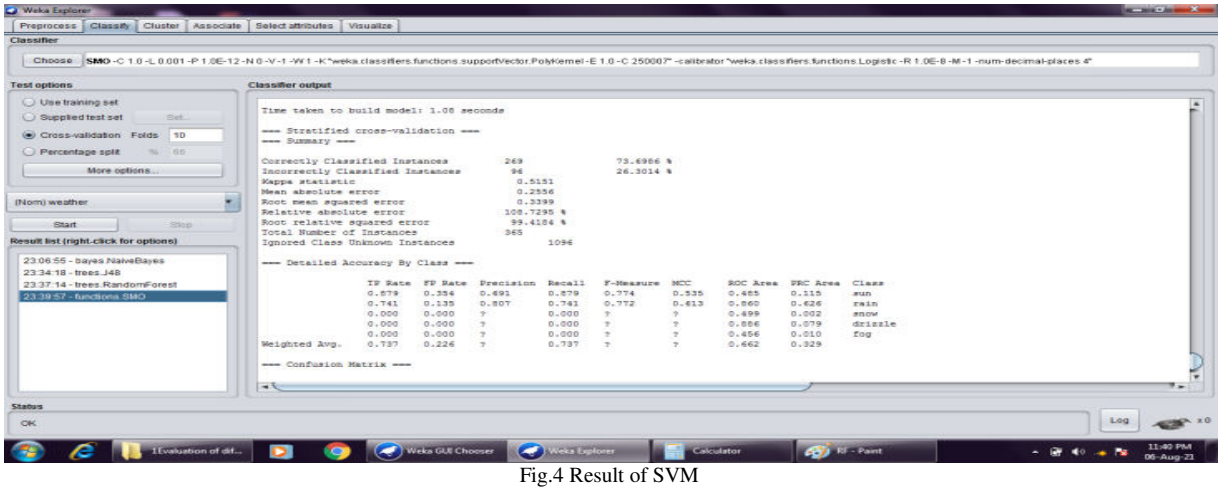

Fig.4 Result of SVM

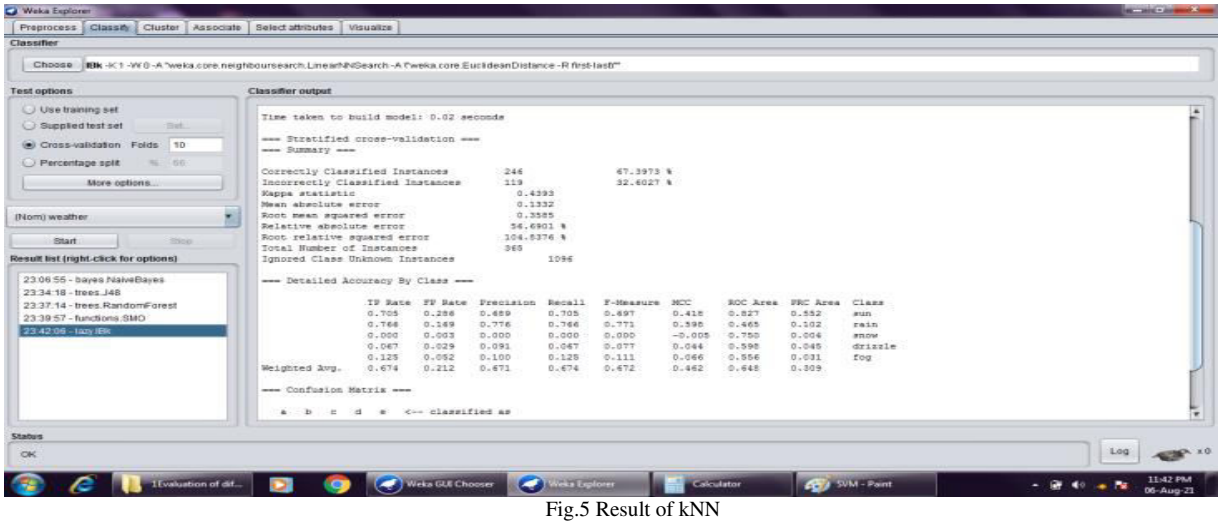

Fig.5 Result of kNN

Now evaluation of all the algorithms are presented in following table, that includes time taken to build up model, correctly classified instances and accuracy.

Table-I Performance Evaluation of Algorithms

Algorithms	Time (seconds)	Correctly classified instances (total 365)	Accuracy (%)
NB	0	283	77.5342
C4.5	0.03	318	87.1233
RF	0.28	322	88.2192
SVM	1.08	269	73.6986
kNN	0.02	246	67.3973

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 8, August 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0908026 |

The following graph shows the performances of the algorithms-

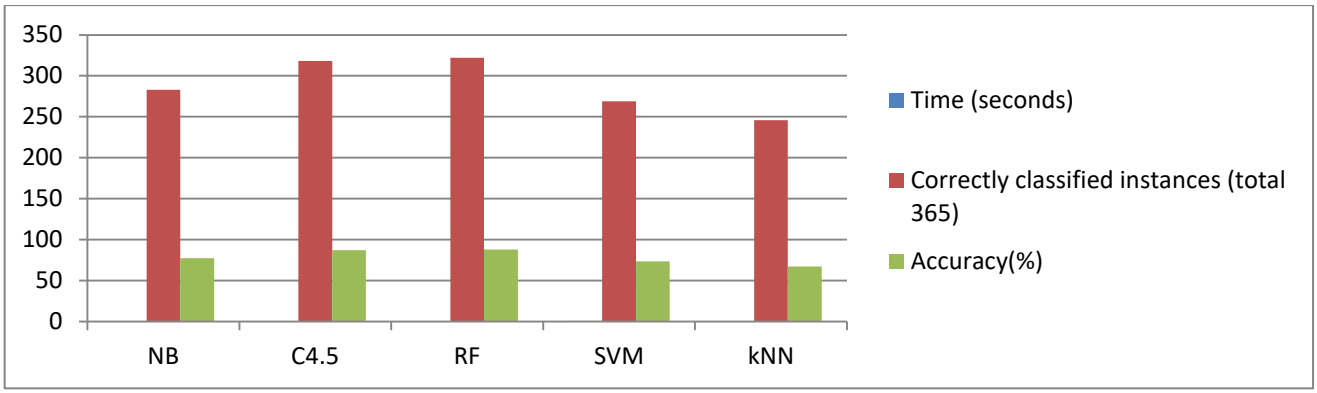

Fig.6-Evaluation Chart

VI. CONCLUSION

Data mining is very popular method for different KDD task. In this paper data mining classification technique is used for weather forecasting.

There are so many facts of classification algorithms we can see in the result and evaluation part like accuracy, precision, recall, specificity, F measure etc. We observe that performance of Random Forest is best of all because its accuracy is highest as well as time taken to build up model is also less.The performance of C4.5 is best after RF. The time of Naïve Bayes is lowest and that is good but accuracy is low. kNN gives worst performance. Performance of SVM is medium between all. So we can say Random Forest is best for weather data prediction.

REFERENCES

[1]. Santos, A. S., Camargo, L. F. R., & Lacerda, D. P. (2020). Evaluation of classification techniques for identifying fake reviews about products and services on the internet. *Gestão & Produção*, 27(4), e4672. https://doi.org/10.1590/0104-530X4672-20

[2]. J. Jayapradha, Kishore Jagan Jothi Kumar, Binti Deka "Educational Data Classification and prediction using Data Mining Algorithms" IJRTE ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

[3]. Mumine "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study" ISSN 1330-3651 (Print), 2019 ISSN 1848-6339 (Online).

[4]. Bayu Adhi Tama and Sunghoon Lim "A Comparative Performance Evaluation of Classification Algorithms for Clinical Decision Support Systems" Mathematics Oct 2020, 8, 1814; doi:10.3390/math8101814..

[5]. Keval Pipalia, Rahul Bhadja "Performance Evaluation of Different Supervised Learning Algorithms for Mobile Price Classification" IJRSET ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VI June 2020.

[6]. Saeed Ngmaldin Bardab, Tarig Mohamed Ahmed, Tarig Abdalkarim Abdalfadil Mohammed "Data mining classification algorithms: An Overview" IASE (2) 2021, Pages: 1-5.

[7]. F. Sherwani, B.S.K.K. Ibrahim , Muhammad Mujtaba Asad "Hybridized classification algorithms for data classification applications: A review" Egyptian Informatics Journal 22 (2021) 185–192.

[8]. Azar Abid Salih1*, Adnan Mohsin Abdulazeez " Evaluation of Classification Algorithms for Intrusion Detection System: A Review" JSCDM VOL. 2 NO. 1 (2021) 31-40.

[9]. Fairoz Q. Kareem1*, Adnan Mohsin Abdulazeez2 and Dathar A. Hasan " Predicting Weather Forecasting State Based on Data Mining Classification Algorithms" Asian Journal of Research in Computer Science 9(3): 13-24, 2021; Article no.AJRCOS.68636 ISSN: 2581-8260.

[10]. Kazheen Ismael Taher1*, Adnan Mohsin Abdulazeez2 and Dilovan Asaad Zebari "Data Mining Classification Algorithms for Analyzing Soil Data" Asian Journal of Research in Computer Science8(2): 17-28, 2021; Article no.AJRCOS.68035 ISSN: 2581-8260.

[11]. J iawai Han and Micheline Kamber Data Mining: Concepts and Techniques, 3rd edition.

[12]. Wikipedia.

[13]. Luis Chaves and Goncalo Marques "Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study" applied sciences MDPI Appi Scie.2021.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details