International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

A Framework for Implementing Machine Learning algorithms using Data sets

J. Vijaya Chandra, G. Ranjith, Arroju Shanthisri, R.Rakshitha, K.Mahesh

Abstract*: The rapid development of cloud computing, big data, machine learning and datamining made information technology and human society to enter new era of technology. Statistical and mathematical analysis on data given a new way of research on prediction and estimation using samples and data sets. Data mining is a mechanism that explores and analyzes many dis-organized or dis-ordered data to obtain potentially useful information and model it based on different algorithms. Machine learning is an iterative process rather than a linear process that requires each step to be revisited as more is learned about the problem. We discussed different machine learning algorithms that can manipulate data and analyses datasets based on best cases for accurate results. Design and Implementation of a framework that is associated with different machine learning algorithms. This paper expounds the definition, model, development stage, classification and commercial application of machine learning, and emphasizes the role of machine learning in data mining by deploying the framework. Therefore, this paper summarizes and analyzes machine learning technology, and discusses the use of machine learning algorithms in data mining. Finally, the mathematical analysis along with results and graphical analysis is given.

***Index Terms*: Machine learning, Data mining, algorithms, datasets and cloud computing.**

I. INTRODUCTION

Cloud based secured data at large volumes with huge storage was required at the modern world, the development of Information Technology in various areas such as big data, cloud computing made data storage and retrieval more secured and can be operated from any place all over the world. The researchers all-over the world started work on large volume data storage and mining that means of retrieval, storage, manipulation and deletion of databases in cloud storage or big data has given rise to a method to manage precious data for further decision making based on different algorithms and datasets. Data mining is also called as knowledge discovery process, it is a process of extraction of useful information and patterns from large volumes of data. knowledge mining from data or data mining is the knowledge-based extraction of data using pattern analysis. Most scholars believe that data mining was first proposed by Fayyad at the Knowledge Discovery Conference in 1995. He believed that data mining was a complex process that automatically or semi-automatically finds effective, meaningful, potentially useful, and easily understood data models from many cloud-based databases. Data mining is a very complex process and requires multi-step iteration. In the process of solving practical problems, scholars have gradually summarized up the process of data mining: The first step is to select the data, usually selecting the appropriate historical data; then, the selected data is preprocessed to eliminate differences and inconsistencies between data. Finally, the data is analyzed, and the interpretable model is obtained, and the generality is verified.

Cloud computing is essentially an extreme form of outsourcing in the delivery of hosted services via the internet, where user search for useful data from a large amount of data. The cloud acts as a virtual server that users can access via the internet on an as need basis. It eliminates the need for companies to host their own servers and purchase expensive software. The process of mining large amount of data logically the goal of this technique is to obtain data using machine learning algorithms for fast retrieval of data without any unwanted or scrap data. Major cost saving for large businesses who managed and store large amounts of data. Cloud computing includes any subscription based or pay-per-use service that extends IT capabilities allowing users to access their stored information remotely. It reduces the need for user to plan for in advance for expansion and loss of data. It removes the need for more and expensive hardware such as memory. Allows users to access data across broad networks on an as needed basis increase the speed and elasticity of the release of certain capabilities. Cloud Computing is not a single technology or architecture it is a platform of computing, networking, security, privacy and storage technologies. Millions of people and organizations around the world adopted cloud computing [1].

The steps involved in logical process of data mining on large volumes of data is to find patterns that are previously unknown, once these patterns are identified they are further used for decision making and statistical analysis for further development of the data retrieval and manipulation. The major steps involved in this process are exploration, pattern identification and deployment [2].

Manuscript published on 30 September 2019.
***Correspondence Author(s)**
J. Vijaya Chandra, Head of the Department & Assistant Professor, Department of Computer Science and Engineering, Warangal Institute of Technology and Science, Warangal, India.
G. Ranjith, Vice Principal and Associate Professor, Department of Computer Science and Engineering, Warangal Institute of Technology and Science, Warangal, India.
Arroju Shanthisri, M.Tech Scholar, Department of Computer Science at Warangal Institute of Technology and Science, Warangal, Telangana, India
R.Rakshitha, M. Tech scholar, Department of Computer science, at Warangal Institute of Technology and Science, Warangal, Telangana, India
K.Mahesh, M. Tech scholar, Department of Computer science, at Warangal Institute of Technology and Science, Warangal, Telangana, India

The Open and distributed service system of cloud computing provides data exploration as a step where the data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined. Where the next step is pattern identification the data is refined and defined after the exploration, where the specific variables and chooses the best prediction and analysis. Finally, the deployment is the step where the implementation takes place,

that is the patterns that are deployed for the desired outcomes [3].

II. RELATED WORK

In cloud computing every component is available through internet, so there are high chances of security breaches that are the risks, threats, vulnerabilities and attacks by intruders and hackers. Social networking became popular over the past decade, cyber attackers using social media for identification of their target. Cloud Storage enables users to store data remotely and retrieve on requirement of the cloud user, without maintaining any software or hardware. Security limiting an organization moving to a cloud. "You can have security and may not have privacy, but you cannot have privacy without security", so the privacy and security are interrelated. The knowledge-based privacy varies widely based among and sometimes within countries, cultures, and jurisdictions. It is shaped by public exceptions and legal interpretations; Privacy rights or obligations are related to the collection, use, disclosure, storage, and destruction of personal data that is personally identifiable information [4].

Data mining is a cross-cutting discipline that needs to combine knowledge and statistical analysis. The main characteristic embodied in the combination of data mining and machine learning is the emphasis on the characteristics and distribution of data. Those features are mainly reflected in the application of machine learning in big data [5].

Machine learning is a learning method that automates the acquisition of knowledge. Machine learning plays an important role in artificial intelligence research. An intelligent system without learning ability cannot be regarded as a real intelligent system, but the intelligent system in the past was generally lack of learning ability. For example, they cannot correct themselves in time when they encounter errors. It does not automatically acquire and discover the required knowledge. Its reasoning is restricted to deduction and lack of induction. Therefore, they can only prove existing facts and theorems, and cannot find new theorems, laws and rules. It does not improve its performance by accumulating experience. With the further development of artificial intelligence, these limitations are becoming more and more prominent. In this situation, machine learning has gradually become one of the core of artificial intelligence. Its application has spread to all over branches of artificial intelligence, such as expert systems, automatic reasoning, natural language understanding, pattern recognition, computer vision, intelligent robots and other fields [6].

According to the machine learning and artificial intelligence viewpoint, learning is the enhancement or improvement of the system's ability in its repeated work based on algorithms and statistical methods. When the system performs the same or comparable tasks at the subsequent time, it will be better or more well-organized. Machine learning is an important way for computers to acquire knowledge and an important indicator of artificial intelligence. It is a discipline that studies how to use computers to simulate or realize human learning activities. It is to study how to make machines obtain new knowledge and skills by identifying and using existing knowledge. It is generally believed that machine learning is a process of acquisition knowledge with a specific purpose. Its internal performance is a process of knowledge growth from unknown to known. Its external performance is the improvement of some performance and adaptability of the system, so that the system can finish the task that could not be completed or better completed [7]. Based on the learning definition we can establish the basic model. The external environment is a collection of external information expressed in some form, which represents the source of external information. Learning is the process of processing external information into knowledge. First, external information is obtained from the environment; and then the information is processed into knowledge and the knowledge is put into the knowledge base. The general principles of execution are stored in the knowledge base. The execution link is the process of using knowledge in the knowledge base to complete a certain task and feeds back some information that obtained in the process of completing the task to guide further study. process of completing the task to guide further study [8]. In the first stage, in the 1950s the main method was to construct the neural network and self organizing learning system, and the learning performance was the feedback adjustment of the transmission signal of the threshold logic unit. In the second stage, in the early 1960s scholars began to study concept-oriented learning, which is symbolic learning. In concept acquisition, the learning system constructs the symbolic representation of concepts by analyzing many positive and negative examples of relevant concepts. In the third stage, in the middle of 1970s, the first symposium on machine learning held in Carnegie Mellon University. It marks machine learning as an independent field of artificial intelligence. In the fourth stage, from the late 1980s to the present, machine learning research has entered into the fields of automation and pattern recognition. Various learning methods began to inherit and began to move from the laboratory to the field of application [9]. The Advanced Persistent Threat is an adversary, that will target an organization, the major aim is to stay for a long-time access and the goal is data extraction. APTs are stealthy, targeted, adaptive and data focused. The APTs are popular and most effective because of their nature that is Persistent. Advanced Persistent Threat Attack has different phases in evaluation cycles majorly intelligence gathering, initial exploitation, command and control, privilege escalation and data exfiltration. Spear phishing has become a very common method used by those launching APTs as an entry point to an organization. To protect the organizations from regular spear phishing mails often email filters are used, the well-designed spear phishes takes only a single user to click on a link and open an attachment for an APT to begin to execute is the first phase on an attack.

Retrieval Number: K12630981119/19©BEIESP
DOI: 10.35940/ijitee.K1263.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Advanced Persistent Attack is an unauthorized access by breaking authentication and security to the cloud and stays for a long time in undetectable stage [10].

III. CLASSIFICATION OF MACHINE LEARNING METHODS

3.1 Rule Induction

Rule induction is to produce a decision tree or a set of decision rules from the training set to classify. The main advantage of rule induction is that it has strong ability to process large data sets and is suitable for classification and predictive tasks. The results are easy to interpret and technically easy to implement.

3.2 Neural networks

the neural network consists of processing nodes like human brain neurons. The input node is connected to the output node through a hidden node to form a multi-layer network structure. The neural network learns through repeated network training on historical sample data. The greatest advantage of neural network is that it can accurately predict complex problems.

3.3 Case reasoning

Each case consists of two parts: problem description and solution to the problem. After asking questions, the system will look for matching cases and solutions. Its advantage is that it can better deal with pollution data and missing data, which is very suitable for many cases.

3.4 Genetic algorithms

Genetic algorithm is a combinatorial optimization method based on biological evolution process. The basic idea is the survival of the fittest and the best or better individual. The operation process includes reproduction, hybridization and mutation. The advantage of genetic algorithm is that it is easy to integrate with other systems.

3.5 Inductive logic programming

Inductive logic programming uses first-level attribute logic to define and describe concepts. Where this technique has strong conceptual description mechanism to express complex relations. First, it defines positive and negative examples and then rank the new examples. Genetic search is based on the Darwinian Theory of survival of fittest. The algorithm carries an inductive learning approach that was initially introduced, this algorithm performs in a way like the genetic models of the natural systems and this is the reason for calling it. It exerts a constant population of individuals to search a sample of space. Every one of the population is evaluated by fitness. Thereafter, new individuals are produced by selecting the outperforming individuals that produces “offspring”. The offspring will retain the features of their parents and generate a population with improved fitness. The process of producing new individuals is done by two significant Genetic Operators i.e., “Crossover” and “Mutation”. Crossover operates by random selection of a point in two parent gene structure and develops two new individuals by exchanging the remaining part of parents [11].

In Knowledge discovery process, Various machine learning algorithms and mechanisms are used such as Classification, Clustering, Regression, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used.

IV. MACHINE LEARNING TASKS IN DATA MINING

In this paper we introduce a framework which uses the data mining. The framework design developed during this paper would provide the numerous data mining techniques, how they can be used for machine learning problems, in what way they work for it. Let us see what the developed framework does individually as a task

A. ***Classification:*** The training data set is used to obtain a classification model. Then, the classification model can automatically divide the data into multiple categories. The existing classification algorithms of machine learning include KNN classification algorithm, naive bayes classification algorithm, decision tree, artificial neural network and support vector Machine, etc [12].

B. ***Regression analysis:*** By analyzing the data and applying statistical methods, the relational expression between variables and variables is obtained. These inherent laws are used to estimate and predict future trends. Regression models are constructed by regression tree, artificial neural network, linear regression and logic regression.

C. ***Association rules****:* There are association rules among transactional data. By mining the relationship between transactional data, frequent itemset can be obtained. Based on this, the probability of certain transactions occurring simultaneously is predicted. Apriori is a classical algorithm for mining association rules.

D. Clustering**:** By using the mining algorithm, multiple data without category labels are aggregated in several different clusters, so that the data objects in the cluster are like each other, and the data objects between clusters are different from each other. K-means is a classical clustering algorithm.

V. ALGORITHMS AND MODELS

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. Classification is a mechanism of categorization; the algorithm then encodes these parameters into a model called a classifier. It uses a specific process for identification of category or class that the data belong. The different Types of Classification models are

a) Bayesian Classification
b) Neural Networks
c) Support Vector Machines (SVM)
d) Classification Based on Associations
e) Random Forest (RF)
f) Decision tree induction

After the classification the next major technique used in the machine learning is clustering, it is identification of similar classes of objects. We can further identify dense and sparse regions by using this mechanism in object space and can discover overall distribution pattern and correlations among data attributes. Its major tasks are attribute selection and data exploration and algorithm evaluation [13]. The different types of clustering methods commonly used are

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

a. Partitioning Methods
b. Hierarchical Agglomerative (divisive) methods
c. Density based methods
d. Grid-based methods
e Model-based methods

Regression adapted for predication where Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. Types of regression methods are

a. Linear Regression
b. Multivariate Linear Regression
c. Nonlinear Regression
d. Multivariate Nonlinear Regression

Association Rules are used to analyze data patterns for a given dataset, is generally very large and a high proportion of the rules, that are usually of little (if any) value. Types of association rule are

a. Multilevel association rule
b. Multidimensional association rule
c. Quantitative association rule

Neural networks have a series of inter-related algorithms that are best at identifying patterns or tendencies on data sets and well suited for prediction or forecasting needs. Types of neural networks

a. Back Propagation

VI. MACHINE LEARNING MODELING

Training a model is the major task in machine learning, here the input data is divided into two datasets that are train dataset and test dataset, in general 70 to 80 percent of the input data and the remaining 20 to 30 percent of test data is considered for the test data. the data in both train and test datasets are of similar in nature as the data division will be taken place randomly, Random Numbers are used to assign data items to the partitions. The model accuracy is given out by total number of correct classifications by the total number of classifications done [15].

The Sensitivity of the machine learning model is measured by the ratio of true positive rate and the combination of true positive and false negative rate. Another measure regarding the accuracy is of confusion matrix which contains correct and incorrect predictions in the form of True positives, False positives, False Negatives and True Negatives.

The percentage of misclassifications is indicated using error rate which is measured as sum of false positive and false negative by sum of true positive, false negative and true negative.

Sometimes the model accuracy can be boosted due to correct prediction of both true positive and true negative, it may happen by coincidence, where their accuracy boosts the mode, where kappa value of a model indicates the adjusted model accuracy [14].

VII. IMPLEMENTATION MECHANISMS AND ANALYSIS

Scikit-learn is a machine learning package that consists of Numpy, which is official known as Numerical Python is a library consists of multi-dimensional array objects. Pandas and mathplotlib a plotting library are used as part of implementation and associated with Scikit-learn along with seaborn is a data visualization library based on matplotlib.

```
import numpy as np
import pandas as pd
import seaborn as sns
%matplotlib inline
from matplotlib import *
import matplotlib.pyplot as plt
```

A dataset spambase is used for testing of machine learning algorithms and implementation of the code, the dataset consists of ham and spam emails, the spam mails are unsolicited or junk emails whereas the ham emails are the legitimate emails. The dataset contains 4601entities, the next step is loading the dataset into pandas using the command **dataset = pd.read_csv('c:\dataset\spambase.csv')**

The next step is organizing data into sets, that is train set and test set, where 70 to 80 percentage of data comes under the train dataset and 20 to 30 percentage of data comes under the test dataset.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2, random_state=10)
```

The implementation of different machine learning algorithms on the dataset is done. We modified the linear regression method and implemented on the data set.

from sklearn.linear_model import linear regression
clf = linearRegression()
clf.fit(X_train, y_train)
clf.predit(X_test)

After implementation and testing the method for the accuracy checking we implemented the following code on the data set.

clf.score(X_test, y_test)

Finally, the percentage of the accuracy is obtained.

VIII. RESULT ANALYSIS

Comparation evaluation of the Regression methods for classification accuracy metrics on corpuses.

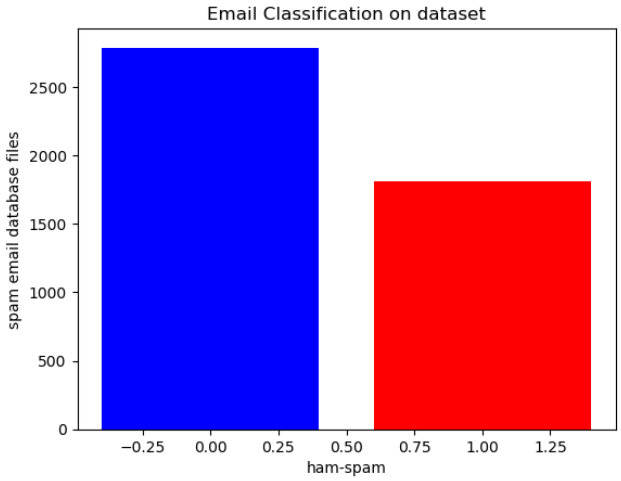
Search Techniques	Metrics	Enron	Spam Assassin	Ling Spam	Spam base
Linear Regression	Accuracy (%)	96.4	87.6	93.2	92.4
Multivariate Linear Regression		97.1	93	94.7	94.9
Nonlinear Regression		96.6	94	96.6	96.2
Multivariate Nonlinear Regression		97.9	94.2	95.4	97.4
Modified Linear Regression		98.4	96	97.9	98.9

IX. GRAPHICAL ANALYSIS

The machine learning methodology using data mining principles using different packages such as Scikit-learn, pandas, Numpy and mathplotlib is used for graphical analysis using the dataset **spambase.csv.**

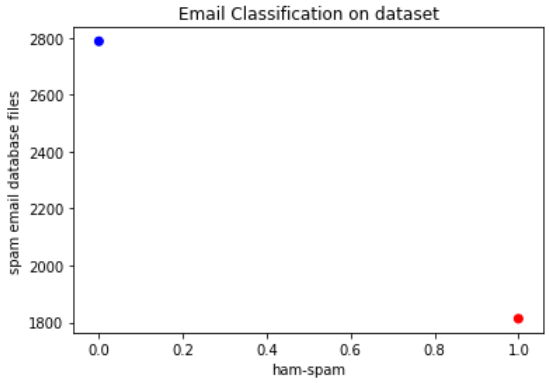
Retrieval Number: K12630981119/19©BEIESP
DOI: 10.35940/ijitee.K1263.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)



Graph 1: Bar Graph of Classification of HAM & SPAM

The Graph based result representation is easier to understand and efficient then representing in tables. In the above graph the classification of ham and spam emails are shown as bar graph representation. In the below graph the Ham emails are classified at 2800 and spam mails are classified around 1800.



Graph 1: Plotter Graph of Classification of HAM & SPAM

X. CONCLUSION

The machine learning mechanism in data mining has been practically implemented in many industries and organizations, including telecommunication industry, banking sector, financial industry, retail industry, insurance industry, and so on. For example, in the financial industry, financial analysts use data mining to build prediction models to identify the patterns that have caused market volatility in history, thereby improving the ability of predicting market volatility. In the retail industry, salespeople can build predictive models through data mining to understand who are most likely to respond to correspondence, thereby increasing sales. When enterprises apply data mining technologies, they should fully understand the advantages and disadvantages of various technologies and methods and select appropriate technologies for specific environments and tasks [16].

REFERENCES

1. A. Nayak, N. K. Sridhar, G. R. Poornima and Shivashankar, "Security issues in cloud computing and its counter measure," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017.
2. J. V. Chandra, N. Challa and S. K. Pasupuleti, "Advanced Persistent Threat defense system using self-destructive mechanism for Cloud Security," 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, 2016, pp. 7-11.
3. J. Vijaya, Chandra Narasimham, M. Ali Hussain, "Data and Information Storage Security from Advanced Persistent Attack in Cloud Computing", International Journal of Applied Engineering Research, 2014.
4. S. R. Gomes et al., "A comparative approach to email classification using Naive Bayes classifier and hidden Markov model," 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, 2017, pp. 482-487.
5. S. R. Guruvayur and R. Suchithra, "A detailed study on machine learning techniques for data mining," *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, Tirunelveli, 2017, pp. 1187-1192.
6. P. Ongsulee, "Artificial intelligence, machine learning and deep learning," *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, Bangkok, 2017, pp. 1-6.
7. M. Xue and C. Zhu, "A Study and Application on MachineLearning of Artificial Intellligence," *2009 International Joint Conference on Artificial Intelligence*, Hainan Island, 2009, pp. 272-274.
8. Q. Zhou and Z. Huang, "A Decision-Making Method Using Knowledge-Based Machine Learning," *2012 International Conference on Computer Science and Electronics Engineering*, Hangzhou, 2012, pp. 616-620.
9. G. R. Ranganathan, Y. Biletskiy and D. MacIsaac, "Machine Learning for Classifying Learning Objects," 2006 Canadian Conference on Electrical and Computer Engineering, Ottawa, Ont., 2006, pp. 280-283.
10. J. V. Chandra, N. Challa and S. K. Pasupuleti, "A practical approach to E-mail spam filters to protect data from advanced persistent threat," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, 2016, pp. 1-5.
11. S. Salehi, A. Selamat and M. Bostanian, "Enhanced genetic algorithm for spam detection in email," 2011 IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, 2011, pp. 594-597.
12. A.Chakrabarty and S. Roy, "An optimized k-NN classifier based on minimum spanning tree for email filtering," 2014 2nd International Conference on Business and Information Management (ICBIM), Durgapur, 2014, pp. 47-52.
13. Chharia and R. K. Gupta, "Email classifier: An ensemble using probability and rules," 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, 2013, pp. 130-136.
14. T. Vyas, P. Prajapati and S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, 2015, pp. 1-7.
15. R. Islam, J. Singh, A. Chonka and W. Zhou, "Multi-classifier Classification of Spam Email on a Ubiquitous Multi-Core Architecture," 2008 IFIP International Conference on Network and Parallel Computing, Shanghai, 2008, pp. 210-217.
16. Adwan Yasin and Addelmunem Abuhasan, "AN INTELLIGENT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION" , Jnternational Jornal of Network Security & Its Applications (IJNSA), Vol.8, No.4,July 2016.

AUTHORS PROFILE

J.Vijaya Chandra, Head of the Department & Assistant Professor, Department of Computer Science and Engineering, Warangal Institute of Technology and Science, Warangal. His Interested Research areas are Cloud Security, Network Security, Intelligence Security and Data Security. Published 15 Research Papers for International Journals. He is Oracle Certified Associate and Member of IEEE and ACM.
Email id: vijayachandra.phd@gmail.com

Dr.Ranjith Gyaderla, Vice Principal and Associate Professor, Department of Computer Science and Engineering, Warangal Institute of Technology and Science, Warangal. His Research Areas are Ontology, Novice, Theory of Computation and Machine Learning. Published 10 Research Papers for International Journals.
Email id: gyaderlaranjith@gmail.com

Arroju Shanthisri is M.Tech Scholar**,** Department of Computer Science at Warangal Institute of Technology and Science, Warangal, Telangana, India. Her Interested areas are Cloud Computing, Machine Learning, Network Security and Artificial Intelligence.
Email id: shanthisriarroju96@gmail.com

Rambathini Rakshitha M. Tech scholar, Department of Computer science, at Warangal Institute of Technology and Science, Warangal, Telangana, India. Her interested areas are network and security, Cloud computing, mobile computing
EMail.id: rakshitha.rambathini@gmail.com

K.Mahesh, M. Tech scholar, Department of Computer science, at Warangal Institute of Technology and Science, Warangal, Telangana, India. Her interested areas are network and security, Cloud computing, mobile computing
EMail.id: kattamahesh04@gmail.com