TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

Loan Default Prediction using Machine learning Techniques

Himanshu Chawla, Bharat Gupta, Govinda.K
SCOPE, VIT, Vellore, India
{chawla08himanshu, bharat.g1999, govindkm}@gmail.com

Article Info
Volume 82
Page Number: 7892 - 7899
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 05 February 2020*

Abstract
Loans are a very fundamental source of any bank's revenue, so they work tirelessly to make sure that they only give loans to customers who will not default on the monthly payments. They pay a lot of attention to this issue and use various ways to detect and predict the default behaviours of their customers. However, a lot of the time, because of human error they may fail to see some key information. The main objective of this work is to automate the process of loan default prediction by using machine learning algorithms like K-Nearest Neighbours, Decision Tree, Support Vector Machine and Logistic Regression to predict defaulters. The accuracy of these methods will also be tested using metrics like Log Loss, Jaccard Similarity Coefficient and F_1 Score. These metrics are compared to determine the accuracy of prediction. This can help banks conserve their manpower and fiscal resources by reducing the number of steps they have to take in order to check if somebody is eligible for a loan.

Keywords:- *Machine Learning, Loan Prediction, Banking, Credit Risk Management, Predictor, Classifiers, Python*

Introduction

The proposed work will deal only with personal loans, but the ideologies used in this project can be applied to other kinds of loans as well. Personal loans grew 20.4% from February 2017 to February 2018. Between the fiscal years 2015 and 2018, unsecured loans had a compounded annual rate of growth as 27% or almost 4 times growth in credit of the bank. The number of individuals with loans rose to 6.54 crores according to an audit done in December 2017. Delinquency rates for the first quarter of last 2018 have remained stable except for a small drop in auto and consumer durable loans. The above statistics have been given by the Reserve Bank of India.

As you can see from the above statistics, in today's world there are a lot of risks involved in taking a loan from a bank. So, to reduce their loss, banks should perform accurate credit risk analysis of an individual before sanctioning his/her loan. If this process is not done well then the loan becomes a bad one as the customer starts defaulting on his loan. Before giving a customer a loan, the bank collects a lot of important details about that individual and then they analyse that information in order to decide whether a person is eligible for a loan or not. A loan defaulter will cause a lot of issue for the bank as they need to dedicate more manpower and time to ensure that they are not facing a loss. Sometimes when people take a loan by mortgaging any property they own, the bank forecloses on that property in order to make up for their loss. So it is imperative for banks to automate at least a small part of their process to make better use of their limited resources.

Most of the banks nowadays use a pen and paper method of loan approval which seems out of step with our now digitised world. It is because of this manual process that they have a slower decision time than what many customers want. However, some banks have automated a part of their work by digitizing the necessary documents and then analysing them. They integrate with credit data services and sources like Digital Matrix Systems. Digital Matrix Systems is an international risk management company that provides strategic risk management solutions. It helps its clients use the power of data to make better decisions for their business. Automating or digitizing their loan

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

sanctioning process ensures that fewer mistakes are made because of human error. Errors like, misplaced documents or bribery will not occur because computers are impartial.
The main focus of this project is the prediction of loan defaulters for personal loans in a financial institution or bank. The proposed system is limited only to predicting which classifier among the given four will be the most accurate for prediction. The most accurate classifier will be decided on the basis of three evaluation metrics. The classifier with the best scores will be the one that is the most accurate for prediction. This project with small modifications can be used in all institutions to predict values. The values of the data set are purely representational and taken only for the sake of the project. Some calculations and assumptions were made considering a proper and realistic model[1-2].

Literature Review
These algorithms have been chosen due to their efficiency and their simplicity. They have a high accuracy and give high percentages of correct predictions.

The most straightforward classifier in all of themachine learning techniques is the K- Nearest Neighbour Classifier. Here, classification is achieved by identifying the nearest neighbours to an example and using those neighbours to ascertain the class of the query. This approach to classification is important today because issues of poor run-time performance are not such a problem these days with the kind of computational power that is available [3-4].

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand [5].

The foundations of SVM have been developed by Vapnik and are gaining popularity in field of machine learning due to many attractive features and promising empirical performance. SVM method does not suffer the limitations of data dimensionality and limited samples [6]

Logistic regression is one of the most commonly used machine learning algorithms for binary classification problems, which are problems with two class values, including predictions such as “this or that,” “yes or no” and “A or B.”. The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular outcomes[7-10].

Proposed Methodology
Data preprocessing
It is imperative to perform data pre-processingprior to analysis in order to get the dataready for analysis. Good data can provide better results hence, data pre-processingis necessary prior to analysisIn data pre-processing, the proposed system performscleaning of data, data imputation, data normalization, and datatransformation. Data cleaningprocess serves to remove the null values and the redundant attributes from a dataset.

Implementation
The proposed system implements machine learning algorithms on the cleaned dataset in order to proceed with the modelling.
The machine learning algorithms being used in this implementation are:

- K- Nearest Neighbours
  K-Nearest Neighbours is a simple algorithm that stores all the available cases and classifies some new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

- Decision Tree
  A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

- Support Vector Machine
  A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side[11-20].

- Logistic Regression
  Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. As more relevant data comes in, the algorithm should get better at predicting classifications within data sets. Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis[21-25].

The metrics being used to test these algorithms are:

- Jaccard Score
  The Jaccard similarity index also called the Jaccard similarity *coefficient* compares members for two different sets to see which of the members are shared and which are distinct. It is a measure of the similarity for the two sets of data, with a range from 0 to 100%.

- F1-Score
  This is a measure of a test's accuracy. It considers the precision *p* and the recall *r* of the test to compute the value of F1-Score: *p* is the number of correct positive results which is divided by the number of all the positive results returned by the classifier, while *r* is the number of correct positive results divided by the number of all the samples that should have been identified as positive. F_1 score is the harmonic mean of the precision and recall, an F_1 score reaches its best value at 1 (perfect precision and recall) and its worst at 0.

- Log Loss
  This measures the performance of a classifier where the prediction input can be a probability value between 0 and 1. The goal of this machine learning model is to minimize the log loss value. An ideal model would have a log loss of 0. The value of log loss increases as the predicted probability differs from the actual label. So predicting a probability of around .012 when the actual observation label is 1 would result in a high log loss[25-30]

Since most bank employees do not have very sound technical skills, it is important to find a solution that they can understand and implement by themselves. The algorithms chosen do exactly that and hence can be used efficiently[31-34].

ALGORITHMS

#K- NEAREST NEIGHBOURS

```
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=4)
print ('Train set:', Xtrain.shape,  ytrain.shape)
print ('Test set:', Xtest.shape,  ytest.shape)
# Modeling
from sklearn.neighbors import KNeighborsClassifier
k = 3 # Taking random value of  k
#Train Model and Predict
kNN_model = KNeighborsClassifier(n_neighbors=k).fit(X_train,y_train)
yhat = kNN_model.predict(X_test)
yhat[0:5]
# Best k
Ks=15
mean_acc=np.zeros((Ks-1))
std_acc=np.zeros((Ks-1))
ConfustionMx=[];
for n in range(1,Ks):
#Train Model and Predict
 kNN_model = KNeighborsClassifier(n_neighbors=n).fit(X_train,y_train)
 yhat = kNN_model.predict(X_test)
 mean_acc[n-1]=np.mean(yhat==y_test);
 std_acc[n-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])
mean_acc
# Building the model again, using k=7
from sklearn.neighbors import KNeighborsClassifier
k = 7
#Train Model and Predict
kNN_model = KNeighborsClassifier(n_neighbors=k).fit(X
train,ytrain)
```

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

Results and Discussion

Data Collection

The proposed system implanted on the historical credit data which iscollected from a kaggle dataset. This dataset contains 346 rows and 12 columns.

Table1. Different metrics

Algorithm	Jaccard Index	F1-Score	LogLoss
KNN	0.88	0.88	0.26
Decision Tree	0.86	0.87	0.31
SVM	0.90	0.90	0.45
Logistic Regression	0.86	0.87	0.49

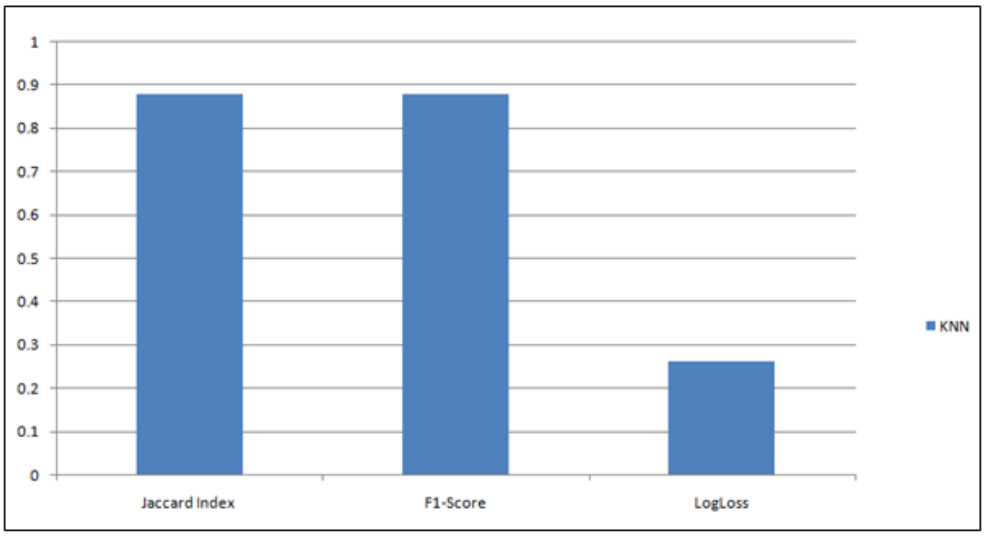

Fig1. K- Nearest Neighbour Approach

In the above graph, the values of the metrics (Jaccard Index, F1-Score and Log Loss) have been plotted for K-Nearest Neighbours.By looking at the above graph we can say that the accuracy values for this algorithm are fairly high compared to algorithms other than Decision Tree, and it has very less log loss.

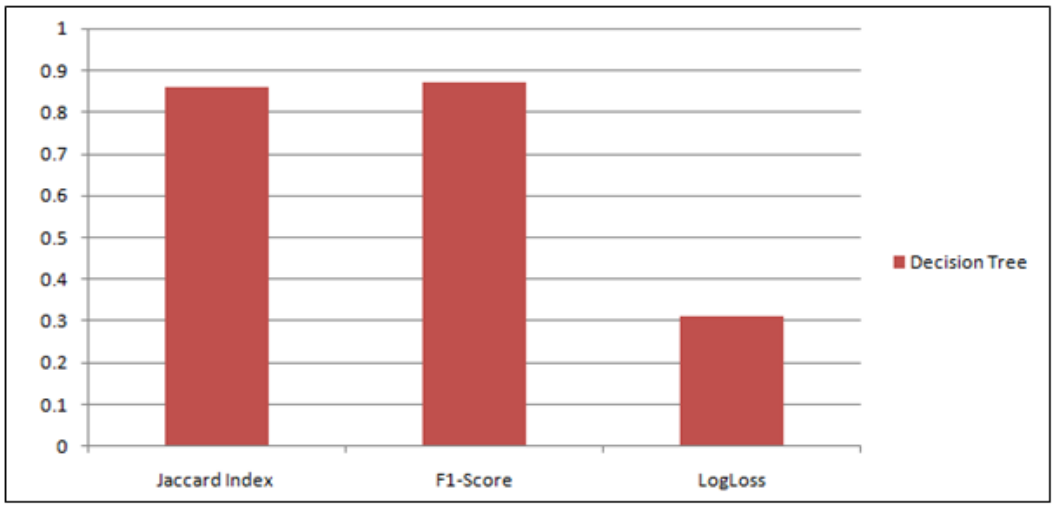

Fig2.Decision Tree

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

In the above graph, the values of the metrics (Jaccard Index, F1-Score and Log Loss) have been plotted for Decision Tree.By looking at the above graph we can say that the accuracy values for this algorithm are very high and it has fairly less log loss compared to metrics other than K-Nearest Neighbours.

Support Vector Machine:

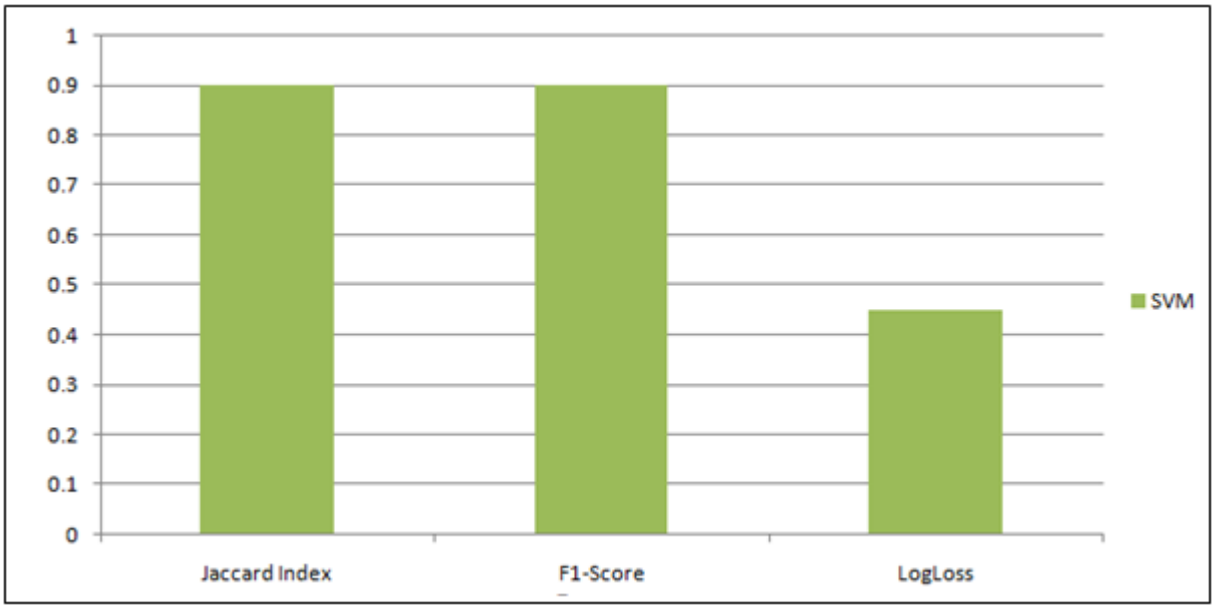

Fig3.Decision Tree

In the above graph, the values of the metrics (Jaccard Index, F1-Score and Log Loss) have been plotted for Support Vector Machine.By looking at the above graph we can say that the accuracy values for this algorithm are the highest and it has a little more log loss compared than above algorithms.

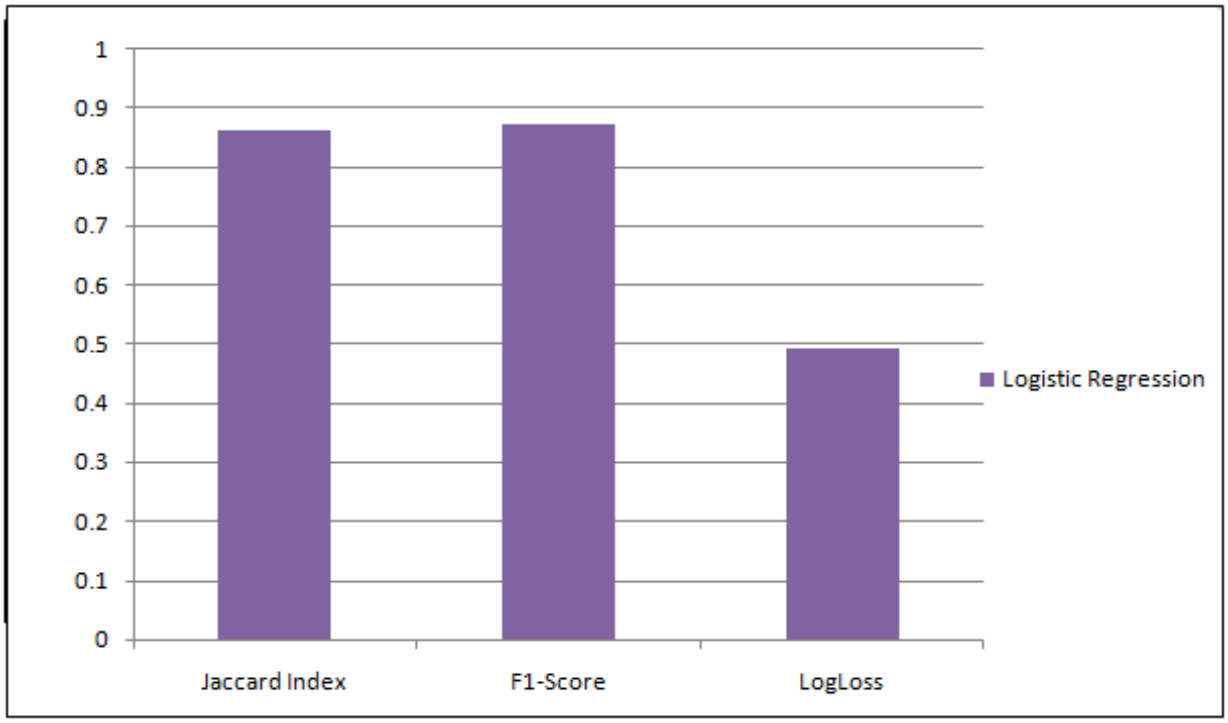

Fig4. Logistic Regression

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

In the above graph, the values of the metrics (Jaccard Index, F1-Score and Log Loss) have been plotted for Logistic Regression. By looking at the above graph we can say that the accuracy values for this algorithm not as high as they were for above algorithms and it has a the most log loss as compared to the above algorithms.

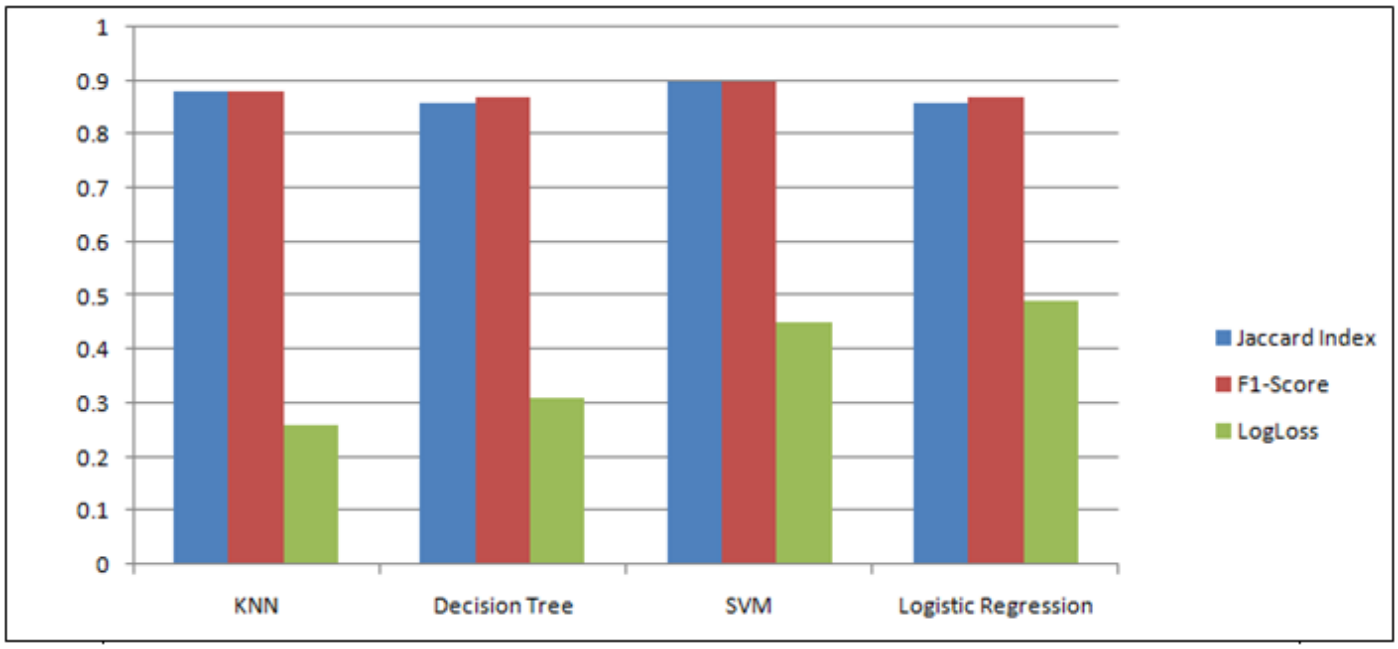

Fig5. Comparative Analysis

From above values and the graph we can say that K- Nearest Neighbours algorithm is the best classifier due to its accuracy values (Jaccard Index F-1 Score) and Log Loss values are best for this particular algorithm.

Conclusion

Based on the above evidence we can say that, the banking process of determining whom to give the loan to, can be automated using machine learning algorithms. From the values obtained we can say that K-nearest neighbours if the best algorithm to use for prediction. This proposed system only goes to show that every aspect of our lives becomes easier when we start automating it. This proposed system can save the bank a lot of monetary resources and precious time by making the initial decisions. This proposed system can be made more accurate by adding more attributes with which modelling can be done which will further increase the accuracy of prediction.All of the innovations of mankind have been used to solve imminent problems, either to protect mankind from their enemies, or help them carry out their day-to-day activities faster and better to sustain health and life. If we aim to automate every aspect of our lives we can achieve a lot more.

References

1. D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. Machine Learning, 6:37–66, 1991.
2. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In Proceedings of 23rd International Conference on Machine Learning (ICML 2006), 2006.
3. Dhilip Kumar V, Vinoth Kumar V, Kandar D, “Data Transmission Between Dedicated Short-Range Communication and WiMAX for Efficient Vehicular Communication” Journal of Computational and Theoretical Nanoscience, Vol.15, No.8, pp.2649-2654, (2018), ISSN: 1546-1963
4. Ruhin Kouser R, Manikandan T,Vinoth Kumar V, “Heart Disease Prediction System Using Artificial Neural Network, Radial Basis Function and Case Based Reasoning ” Journal of Computational and Theoretical Nanoscience,Vol.15,No(9/10),pp.2810-2817,(2018), ISSN: 1546-1963
5. V Vinoth Kumar, K. S. Arvind, S. Uma Maheswaran, Suganya K.S, “Hierarchal Trust Certificate Distribution using Distributed CA in MANET” International Journal of Innovative Technology and Exploring Engineering, (2019), Vol.08, Issue.10, pp.2521-2524

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

6. Shalini A, Jayasuruthi L, Vinoth Kumar V, "Voice Recognition Robot Control using Android Device" Journal of Computational and Theoretical Nanoscience,2018, ISSN: 1546-1963
7. Umamaheswaran, S., Lakshmanan, R., Vinothkumar, V. et al. New and robust composite micro structure descriptor (CMSD) for CBIR. International Journal of Speech Technology (2019) doi:10.1007/s10772-019-09663-0 (Springer)
8. Dhilip Kumar V, Vinoth Kumar V, Kandar D, "Data Transmission Between Dedicated Short-Range Communication and WiMAX for Efficient Vehicular Communication" Journal of Computational and Theoretical Nanoscience, Vol.15, No.8, pp.2649-2654, (2018), ISSN: 1546-1963
9. Karthikeyan T, Karthik Sekaran, Vinoth kumar V, Balajee J M, "Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques" International Journal of Web Portals (IJWP), 11(2), pp.41-52, (2019)
10. E. J. Keogh, S. Lonardi, and C. Ratanamahatana. Towards parameter-free data mining. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, KDD, pages 206–215. ACM, 2004.
11. Basu, S., Kannayaram, G., Ramasubbareddy, S., & Venkatasubbaiah, C. (2019). Improved Genetic Algorithm for Monitoring of Virtual Machines in Cloud Environment. In *Smart Intelligent Computing and Applications* (pp. 319-326). Springer, Singapore.
12. Somula, R., & Sasikala, R. (2018). Round robin with load degree: An algorithm for optimal cloudlet discovery in mobile cloud computing. *Scalable Computing: Practice and Experience*, *19*(1), 39-52.
13. Somula, R., Anilkumar, C., Venkatesh, B., Karrothu, A., Kumar, C. P., & Sasikala, R. (2019). Cloudlet services for healthcare applications in mobile cloud computing. In *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology* (pp. 535-543). Springer, Singapore.
14. Somula, R. S., & Sasikala, R. (2018). A survey on mobile cloud computing: mobile computing+ cloud computing (MCC= MC+ CC). *Scalable Computing: Practice and Experience*, *19*(4), 309-337.
15. Somula, R., & Sasikala, R. (2019). A load and distance aware cloudlet selection strategy in multi-cloudlet environment. *International Journal of Grid and High Performance Computing (IJGHPC)*, *11*(2), 85-102.
16. Somula, R., & Sasikala, R. (2019). A honey bee inspired cloudlet selection for resource allocation. In *Smart Intelligent Computing and Applications* (pp. 335-343). Springer, Singapore.
17. Nalluri, S., Ramasubbareddy, S., & Kannayaram, G. (2019). Weather Prediction Using Clustering Strategies in Machine Learning. *Journal of Computational and Theoretical Nanoscience*, *16*(5-6), 1977-1981.
18. Sahoo, K. S., Tiwary, M., Mishra, P., Reddy, S. R. S., Balusamy, B., & Gandomi, A. H. (2019). Improving End-Users Utility in Software-Defined Wide Area Network Systems. *IEEE Transactions on Network and Service Management*.
19. Sahoo, K. S., Tiwary, M., Sahoo, B., Mishra, B. K., RamaSubbaReddy, S., & Luhach, A. K. (2019). RTSM: response time optimisation during switch migration in software-defined wide area network. *IET Wireless Sensor Systems*.
20. Somula, R., Kumar, K. D., Aravindharamanan, S., & Govinda, K. (2020). Twitter Sentiment Analysis Based on US Presidential Election 2016. In *Smart Intelligent Computing and Applications* (pp. 363-373). Springer, Singapore.
21. Sai, K. B. K., Subbareddy, S. R., & Luhach, A. K. (2019). IOT based Air Quality Monitoring System Using MQ135 and MQ7 with Machine Learning Analysis. *Scalable Computing: Practice and Experience*, *20*(4), 599-606.
22. Somula, R., Narayana, Y., Nalluri, S., Chunduru, A., & Sree, K. V. (2019). POUPR: properly utilizing user-provided recourses for energy saving in mobile cloud computing. In *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology* (pp. 585-595). Springer, Singapore.
23. Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017, October). Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1-5). IEEE.

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 7892 - 7899

24. Somula, R., & Sasikala, R. (2019). A research review on energy consumption of different frameworks in mobile cloud computing. In *Innovations in Computer Science and Engineering* (pp. 129-142). Springer, Singapore.
25. Kumar, I. P., Sambangi, S., Somukoa, R., Nalluri, S., & Govinda, K. (2020). Server Security in Cloud Computing Using Block-Chaining Technique. In *Data Engineering and Communication Technology* (pp. 913-920). Springer, Singapore.
26. Kumar, I. P., Gopal, V. H., Ramasubbareddy, S., Nalluri, S., & Govinda, K. (2020). Dominant Color Palette Extraction by K-Means Clustering Algorithm and Reconstruction of Image. In *Data Engineering and Communication Technology* (pp. 921-929). Springer, Singapore.
27. Nalluri, S., Saraswathi, R. V., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Chronic Heart Disease Prediction Using Data Mining Techniques. In *Data Engineering and Communication Technology* (pp. 903-912). Springer, Singapore.
28. Krishna, A. V., Ramasubbareddy, S., & Govinda, K. (2020). Task Scheduling Based on Hybrid Algorithm for Cloud Computing. In *International Conference on Intelligent Computing and Smart Communication 2019* (pp. 415-421). Springer, Singapore.
29. Srinivas, T. A. S., Ramasubbareddy, S., Govinda, K., & Manivannan, S. S. (2020). Web Image Authentication Using Embedding Invisible Watermarking. In *International Conference on Intelligent Computing and Smart Communication 2019* (pp. 207-218). Springer, Singapore.
30. Krishna, A. V., Ramasubbareddy, S., & Govinda, K. (2020). A Unified Platform for Crisis Mapping Using Web Enabled Crowdsourcing Powered by Knowledge Management. In *International Conference on Intelligent Computing and Smart Communication 2019* (pp. 195-205). Springer, Singapore.
31. Saraswathi, R. V., Nalluri, S., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Brilliant Corp Yield Prediction Utilizing Internet of Things. In *Data Engineering and Communication Technology* (pp. 893-902). Springer, Singapore.
32. Kalyani, D., Ramasubbareddy, S., Govinda, K., & Kumar, V. (2020). Location-Based Proactive Handoff Mechanism in Mobile Ad Hoc Network. In *International Conference on Intelligent Computing and Smart Communication 2019* (pp. 85-94). Springer, Singapore.
33. Bhukya, K. A., Ramasubbareddy, S., Govinda, K., & Srinivas, T. A. S. (2020). Adaptive Mechanism for Smart Street Lighting System. In *Smart Intelligent Computing and Applications* (pp. 69-76). Springer, Singapore.
34. Srinivas, T. A. S., Somula, R., & Govinda, K. (2020). Privacy and Security in Aadhaar. In *Smart Intelligent Computing and Applications* (pp. 405-410). Springer, Singapore.