International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Prediction of Top Tourist Attraction Spots using Learning Algorithms

Sagar Gupta, Jenila Livingston L.M., Agnel Livingston L.G.X.

Abstract: ***Dealing with the growing amount of user posted content like preferences, responses, comments, past experiences and beliefs spread through social media is a vital but challenging task. Being applied in several domains, recommender systems are used to find solutions and suggestions based on users interests including tourism-related opinion detection and tourist-attraction spot identification. Tourists can access and analyze this information for making decisions and predicting best tourist places. This study aims to predict tourist attraction spots and their related information by analyzing the data from social media (Facebook, Twitter etc.) which in turns help the tourist industry by deliberating what kind of attractions tourists can have and how to obtain their preferences. For this purpose four algorithms such as Kernel Density Estimation, K- Nearest Neighbor, Random forest and XG Boost have been used. The findings revealed that XG Boost yields better results in terms of accuracy than other three algorithms.***

Index Terms: ***Tourism, Social media, Random Forest, Kernel Density Estimation, K-Nearest Neighbour, XG Boost, Classification algorithm***

I. INTRODUCTION

It is often necessary for travelers to consult with experts, natives, or friends about which tourism attractions to visit at their desired destinations (e.g., where to visit, where to stop, and how to get there, Customs and Arrival rules, advice, and so on).

Exchanging travel-related information between travelers using the Internet and online social networks has become more important and effective. Different tourist spots were grouped and categorized according to their profiles using ranking method [1]. People have access to incredible extents of travel information, but finding and/or classifying the most relevant information is problematic.

Manuscript published on 30 September 2019
* Correspondence Author
Sagar Gupta*, M. Tech, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India -600127.
Jenila Livingston L.M., Associate Professor, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India -600127
Agnel Livingston L.G.X., Assistant Professor, St. Xavier's Catholic College of Engineering, Chunkankadai, Nagercoil, Kanyakumari District, Tamil Nadu, India -629003

People used to share text, images, audio and video of their events / places visited. Facebook users mostly share their data (nearly 57%) in image format but at the same time most of the users post textual views about the places they visited in addition to sharing their photographs. Analyzing the social media is getting more popular and useful for advertising and predicting the top influential users and ranking most important data. Page ranking algorithms and its revisions are used in most of the cases for predicting top n users/ websites/ places in a social network. These types of study are extremely helpful to promote product/campaign or protect unwanted rumors among its users. Raamakirtinan and Jenila [2] used Sentiment Weighted Page Ranking Algorithm for predicting top influential users in Facebook social media. Usually, countries and cities invest a huge amount of money in planning and preparation in order to welcome tourists [3]. This paper discusses four algorithms for predicting the top tourist spots. Furthermore, thorough comparison of four methods in terms of accuracy has been made for recommending top 'n' touristic spots.

II. LITERATURE SURVEY

Social media data is used for analysis and prediction of traffic. Sridevi et. al. [4] has conducted the study using random forest algorithm for analyzing the Twitter data using Twitter API or Twitter archive, and then tweets ware classified based on location. It also dealt the number of active participants in each place based on the number of tweets.

Tourism plays a vital role in local economies. It is necessary to inspire foreign tourists to stimulate local economies. Maeda et. al. [5] has conducted the study to understand the characteristics of foreign tourists expect of areas near tourist attractions compared with what domestic tourists expect. In this study, a tourist destination is defined small areas focusing the historic sites, theme parks, hotels, and restaurants. The proposed method successfully extracted the locations of tourist destinations and characterizes those locations based on the points of interest in the neighborhood.

Baraglia et. al. [6] used supervised learning techniques such as Gradient Boosted Regression Trees and Ranking SVM for predicting the "next" geographical position of a tourist. The learning is done on the basis of an object space represented by a 68 dimension feature vector, specifically designed for tourism related data. In this study, four prediction algorithms viz., K-Nearest Neighbor model, Random Forest model, XG Boost model and Kernel Density Estimation model has been used for predicting the top tourist spots.

The details of four algorithms and their implementation for prediction are discussed in further sections.

III. PROPOSED METHOD

The modules of the proposed system are illustrated in Fig.1. Dataset including Place-id, latitude, longitude and time are given as an input to the prediction algorithms "Model A - K-Nearest Neighbor model", "Model B - Random Forest model", "Model C- XG Boost model" and "Model D - Kernel Density Estimation (Non-Parametric model)".

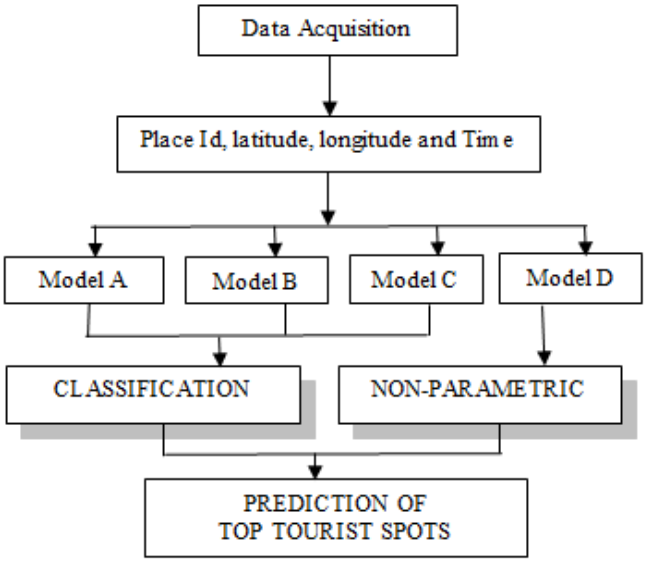


Fig. 1. The modules of a Proposed System

The system is designed as a stand-alone terminal offering users by predicting the top tourist spots based on classification results. All data is stored in a Microsoft Excel spreadsheet and is retrieved and operated by Microsoft-Excel as csv result.

Data Collection

The dataset was collected mainly from Google and its size after removing invalid samples is n=29118021. The data information is described as follows:

```
In [2]: train.info()
        print(train)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29118021 entries, 0 to 29118020
Data columns (total 6 columns):
row_id      int64
x           float64
y           float64
accuracy    int64
time        int64
place_id    int64
dtypes: float64(2), int64(4)
memory usage: 1.3 GB
          row_id       x       y  accuracy    time    place_id
0              0  0.7941  9.0809        54  470702  8523065625
1              1  5.9567  4.7968        13  186555  1757726713
2              2  8.3078  7.0407        74  322648  1137537235
3              3  7.3665  2.5165        65  704587  6567393236
4              4  4.0961  1.1307        31  472130  7440663949
5              5  3.8099  1.9586        75  178065  6289802927
6              6  6.3336  4.3720        13  666829  9931249544
7              7  5.7409  6.7697        85  369002  5662813655
8              8  4.3114  6.9410         3  166384  8471780938
9              9  6.3414  0.0758        65  400060  1253803156
10            10  2.0173  4.8627         6   21353  8684462954
```

Fig. 2. Dataset details

A. K- Nearest Neighbor

K Nearest Neighbor (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already at the beginning of 1970's. It compares the check-in input with other data set that is closest in terms of their Euclidean distance to the check-ins. Here all the processes happen at the time of prediction therefore, it takes more computation time.

B. Random Forest

Random forest and random decision forests are ensemble learning methods for classification and regression. It is a bagging method that operates by constructing an assemblage of decision trees at training time and outputting the class that is the mode of the classes for classification or mean prediction for regression of the individual trees. It consists of a large number of individual decision trees that operate as an ensemble as given in Fig 3.

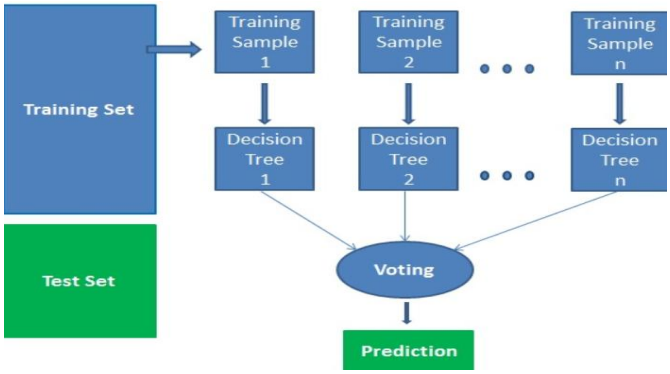


Fig. 3. Random Forest

$$f = \sum_{i=b}^{B} f^{b}(x')$$

Where f is the total function, b is the bagging, f^b is the bagging function, x' is the train sample data

C. XG Boost

XG Boost is based on boosting method which reduces the error of training data by taking the probability of misclassified point every time and gradient help to find the loss every tie so it learns from the previous model and improves the accuracy. It is an ensemble of additive model as shown in Fig 4 or base learners that is composed of several base learners by choosing a function that minimizes the overall loss. So in this system it starts with simple regression on check-in data which give some prediction after calculating the error and generate new prediction function by adding the error with the previous function.

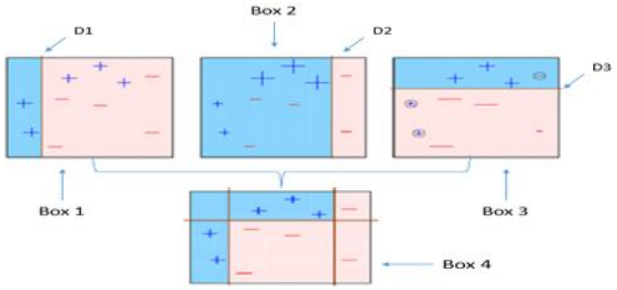


Fig. 4. XG Boost

$$-f_{1(x)} = f_{0(x)} + h_{1(x)}$$
$$-f_{2(x)} = f_{1(x)} + h_{2(x)}$$
$$-f_{3(x)} = f_{2(x)} + h_{3(x)}$$
$$-f_{4(x)} = f_{3(x)} + h_{4(x)}$$
$$-f_{n(x)} = f_{n(x)} + h_{n(x)}$$

$f_0(x)$ - It is a simple linear regression function
$h_1(x)$ - hypothesis to fix the error
$f_1(x)$-It is a new function after fixing the misclassified points

D. Kernel Density Estimation

Kernel Density Estimation is a statistical and non-parametric algorithm that estimates the probability of kernel density function where the kernel is a function which estimates the check-in density of each grid and to find out the most accurate grid structure, needs to cross-validate it by finding the best bandwidth parameter. It mainly explored spatiotemporal data to estimate the check-in density where Gaussian kernel function is used to calculate the spatiotemporal kernel density of each trajectory.

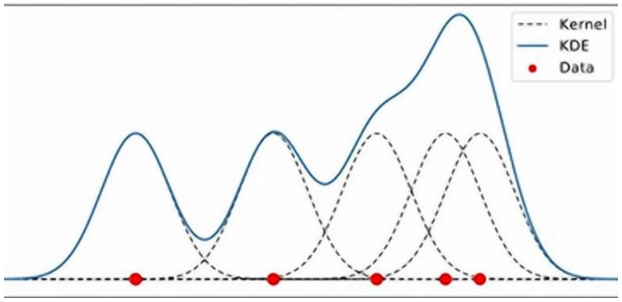


Fig. 5. Kernel Density Method

$$f(l'|l,h) = \frac{1}{n}\sum_{i=1}^{n} K(l' - l_i)$$

Where k ( ) is Gaussian kernel function, L (l1, l2, l3, …. ln) is the given set of locations, l' is 2D dimensional record in the form of (latitude, longitude) of each record, if L points are far away from l' similarity value then F (l'| L, h) will be large or if L points are close to similarity value then F (l'| L, h) will be small.

$$k(l'|l_i) = \frac{1}{2\pi h}\ e^{(-(\frac{l'-l_i}{2h^2}))}$$

IV. RESULTS AND DISCUSSION

The above said four models have been implemented for a given dataset and the results obtained are illustrated in Fig 6, Fig 7, Fig 8 and Fig 9.

```
Starting...
Feature engineering on train
Feature engineering on validation
Data prepared in: 0:11:22.113811
Row 0 completed in: 0:03:08.207733
Row 1 completed in: 0:00:42.875825
Row 2 completed in: 0:00:19.587004
Row 3 completed in: 0:00:19.532783
Row 4 completed in: 0:00:29.318206
Row 5 completed in: 0:00:22.548266
Row 6 completed in: 0:00:19.419971
Row 7 completed in: 0:00:20.197026
Row 8 completed in: 0:00:19.765706
Row 9 completed in: 0:00:19.959122
Row 10 completed in: 0:00:22.072885
Row 11 completed in: 0:00:19.652408
Row 12 completed in: 0:00:20.226352
Row 13 completed in: 0:00:20.080604
Row 14 completed in: 0:00:21.587254
Row 15 completed in: 0:00:20.745577
Row 16 completed in: 0:00:19.874945
Row 17 completed in: 0:00:20.758258
Row 18 completed in: 0:00:18.048371
Row 19 completed in: 0:00:18.068012
Predictions made in: 0:21:29.947781
Final score: 0.53710806
Task completed in: 0:21:32.956006
```

Fig. 6. Prediction of Top Tourist spots using KNN

The accuracy (final score) of KNN algorithm is 54% whereas Random forest algorithm gives 62% accuracy.

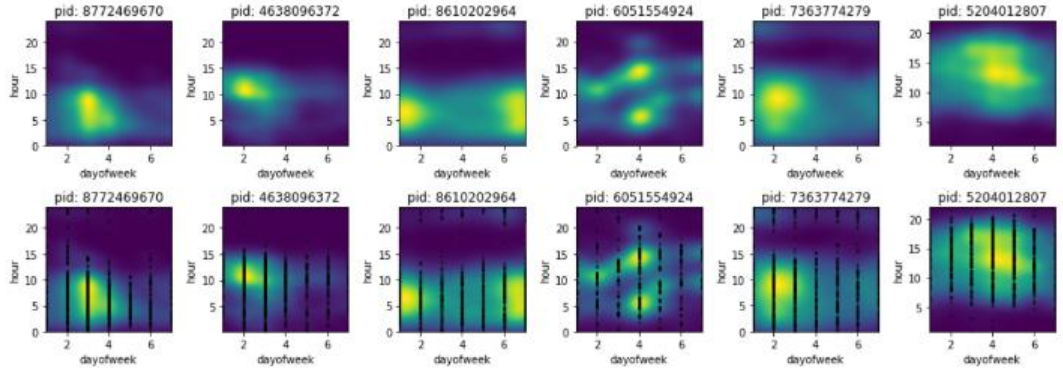
```
9585176434      0.00      0.00      0.00         1
9637829129      0.00      0.00      0.00         1
9714559066      0.00      0.00      0.00         1
9720355171      0.00      0.00      0.00         1
9756528108      0.00      0.00      0.00         2
9766354664      0.00      0.00      0.00         1
9768374581      0.46      0.70      0.56        56
9802999183      0.00      0.00      0.00         1
9854879010      0.00      0.00      0.00         1
9856360779      0.00      0.00      0.00         2
9992656249      0.00      0.00      0.00         4

   micro avg      0.53      0.53      0.53      2819
   macro avg      0.11      0.11      0.11      2819
weighted avg      0.45      0.53      0.48      2819

Iter:16/16 GridX:5.20-5.40 GridY:5.20-5.40
Train - Time:6.647 Size:9803 || Test - Time:2.234 Size:2819 Map@3:0.62268
```

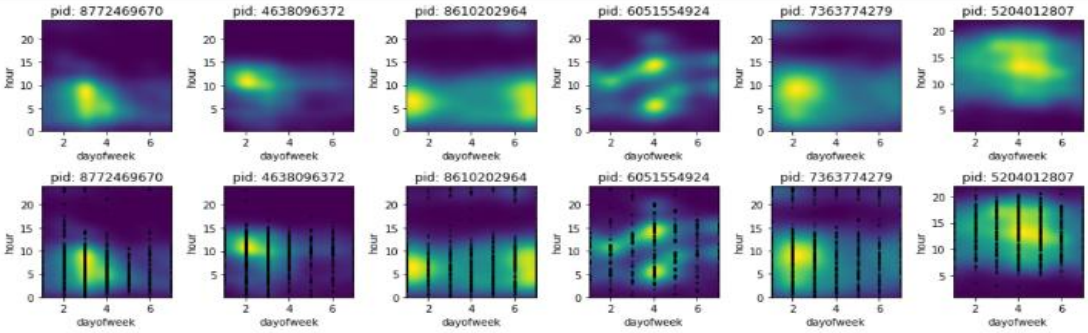
Fig 7: Prediction of Top Tourist spots using Random Forest algorithm

The statistical estimated accuracy of KDE algorithm is 57% which is shown in Fig 8.



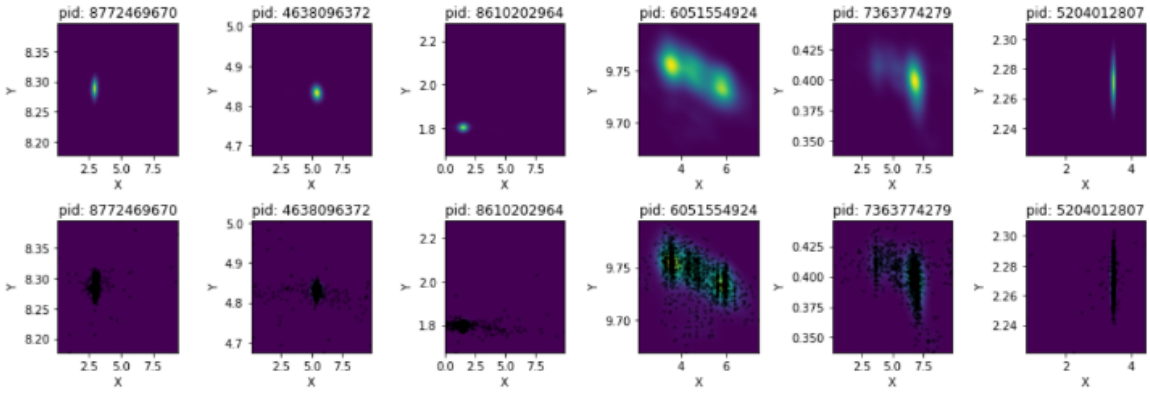

```
In [85]: from sklearn.neighbors.kde import KernelDensity
         import numpy as np
         kde = KernelDensity(kernel='gaussian', bandwidth=0.3).fit(z)
         z1=kde.score_samples(z)
         print("score",sum(z1)/len(z1))

score 56.980210354873506
```




```
In [85]: from sklearn.neighbors.kde import KernelDensity
         import numpy as np
         kde = KernelDensity(kernel='gaussian', bandwidth=0.3).fit(z)
         z1=kde.score_samples(z)
         print("score",sum(z1)/len(z1))

score 56.980210354873506
```




```
In [73]: from sklearn.neighbors.kde import KernelDensity
         import numpy as np
         kde = KernelDensity(kernel='gaussian', bandwidth=0.3).fit(z)
         z1=kde.score_samples(z)
         print("score",sum(z1)/len(z1))

score 57.00675127331699
```

Fig. 8. Prediction of Top Tourist spots using KDE algorithm

```
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 32 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 32 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 18 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 36 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 30 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 24 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 24 extra node
s, 0 pruned nodes, max_depth=6
[20:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning end, 1 roots, 36 extra node
s, 0 pruned nodes, max_depth=6
Train Cross Validated MAP@3: 0.9506
MAP@3 CV in bins [0.9642087, 0.9506003]
```

Fig 9. Prediction of Top Tourist spots using XG Boost

The accuracy of prediction using four learning algorithms viz., KNN, KDE, Random forest, and XG Boost is tabulated in Table1. The results shows XG Boost produces 95% accuracy comparing to other algorithms and this model is best suitable for prediction of tourist attraction places.

Table1: Results of Learning Algorithms

S.no	Methods Used	Accuracy
1.	**KNN**	0.54
2.	**KDE**	0.57
3.	**Random Forest**	0.62
4.	**XG Boost**	0.95

IV. CONCLUSION

This research work focuses suggesting top tourist attraction spots by evaluating opinionated sentences from social media users. We used four machine learning algorithms such as KNN, KDE, Random forest, and XG Boost to train and test the classifiers for tourism-related opinion mining. The accuracy scores of tourist attract spot detection were 54%, 57%, 62% and 95%, respectively and our results shows XG Boost model produces more accuracy.

REFERENCES

1. Chareyron G., Branchet B. and Jacquot S., 2015. “A new area tourist ranking method”, *IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, pp.2930-2932.
2. Raamakirtinan, S. and Jenila Livingston L. M.. 2016. “Identifying Influential Users in Facebook - A Sentiment Based Approach”, *Indian Journal of Science and Technology*, Vol 9(10), pp. 1-9. DOI: 10.17485/ijst/2016/v9i10/86735.
3. Khatibi, A., Belem, F., Silva, A.P., Shasha, D. and Goncalves, M.A., 2018, June. “Improving tourism prediction models using climate and social media data: A fine-grained approach”. *In Twelfth International AAAI Conference on Web and Social Media*. pp.636-639.
4. Sridevi K., Ganesan T., Samrat B V S, Srihari S. 2019. “Traffic Analysis by Using Random Forest Algorithm Considering Social Media Platforms”, *International Journal of Recent Technology and Engineering (IJRTE)*, ISSN: 2277-3878, Volume-7, Issue-6S. pp.620-625
5. Maeda, T., Yoshida, M., Toriumi, F. and Ohashi, H., 2018. “Extraction of tourist destinations and comparative analysis of preferences between foreign tourists and domestic tourists on the basis of geotagged social media data. ISPRS”. *International Journal of Geo-Information*, 7(3), p.99.
6. Baraglia, R., Muntean, C., Nardini, Franco M. & Silvestri, F. 2013. “LearNext: learning to predict tourists movements”. pp.751-756. 10.1145/2505515.2505656.

AUTHORS PROFILE

Mr. Sagar Gupta is with M. Tech, specialization in Big Data, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India - 600127. (e-mail: sagargupta.2018@vit.ac.in). He is very much interested in learning new concepts explore and implement novel ideas. His area of interest for research and publication are in Big data analytics, Data Science and Machine Learning.

Dr. Jenila Livingston L. M., Associate Professor, is with School of Computing Science and Engineering, VIT, Chennai, TN 600127, INDIA. (e-mail: jenila.lm@vit.ac.in). She has completed her PhD in Faculty of Engineering from National Institute of Technical Teachers’ Training and Research (NITTTR), Government of India, Chennai and Master’s Degree in Computer Science and Engineering from Anna University, India. She has nearly 15 years of experience in Teaching and Research and keenly interested in the areas of eLearning, Engineering Education, Artificial Intelligence, Soft Computing, Data Analytics, Internet & Web Programming, Data Base Systems and Data Structures & Algorithms.

Prof. Agnel Livingston L. G. X., Assistant Professor, is with Department of Computer Science and Engineering, St. Xavier’s Catholic College of Engineering, Chunkankadai, Kanyakumar District, TN 600127, INDIA. (e-mail: agnellivingston@yahoo.com). He has completed Master’s Degree in Computer Science and Engineering from Anna University, Chennai, India. He has nearly 10 years of experience in Teaching and Research and keenly interested in the areas of Image Processing, Computer Network, Computer Architecture, Data Analytics, Data Structures & Algorithms and Data Base Systems.