International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

# Descriptive Analytics on Crime in India using Clustering Techniques

**J. Vimala Devi, Kavitha K S**

***Abstract*: Understanding the perspectives of crime happenings by exploiting crime data helps early detection and prevention of crimes. Ruling Government and Policing systems are aware the importance of realising the changing aspects of crime. As technology is advanced, there are many ways to comprehend the trends and the patterns of crime activities. The paper presents a hybrid model using AGNES and K-means clustering algorithms to focus on different views and representation of crime in India and aims to recognize the crime types that cluster the regions of India. Accuracy of model is measured using log loss and the patterns and trends in crime are presented.***

***Keywords: Unsupervised learning, Clustering techniques, Machine learning, Crime data analytics***

## I. INTRODUCTION

India is one of the densely populated and developing country. Stable rise in urbanization, high population and poverty, lack of jobs and lack of Education for all are the factors that led to increase in crime rate, every year. The increase in crime rate affects the economic growth and repute of a country and is also a major threat to the citizens to carry out their day to day activities.

In this paper, we focus on providing crime data visualization in different perspectives and this paper helps policing systems in understanding crime patterns effectively and in a systematic and smarter way. There are many statistical analysis techniques and machine learning algorithms exist for our disposal. Depending on the dataset, objectives and end users of this system, the techniques and methodology used are probability density functions, cumulative density function, heat maps, word clouds and clustering algorithms.

This paper answers for the following questions:

- Is there any change in the crime rate year wise?
- Is there any possibility of clustering the locations based on the number of crimes? If so, does this data visualisation show any clusters?
- What are the other ways of visualising the data?

**Revised Manuscript Received on January 30, 2020.**
***** Correspondence Author
**J Vimala Devi***, Department of Computer Science and Engineering, Visveswaraya Technological University, Bangalore, India. Email: vimalajana@gmail.com
**Dr Kavitha K S,** Department of Computer Science and Engineering, Visveswaraya Technological University, Bangalore, India. Email: drkavitha2015@gmail.com



## II. LITERATURE STUDIES

**2.1 Crime forecasting:** The objective of this paper [1] is to foretell crime rate in India every year using the Time Series Models such as Auto-Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing. Source of data is from the National Crime Record Bureau of India. While building the predictive model, data is divided into training and test data.. Accuracy of the model is examined and from Hypothesis testing, it is evident that the forecast values are within 95% confidence interval of the test data. This paper concludes that crime forecasting can be done by building time series model.

**2.2 Crime analysis using data mining:** This paper [2] is devoted to present the perspectives of Using Data mining methods in crime analysis. The Data mining methods discussed in this paper, suggest the process of developing and employing proactive policing strategies for the prevention and investigation of crimes. This paper lists the data mining tools for the efficacy of data analysis required by law-enforcement agencies by designing intelligent tools. It also outlines the models and technologies that automate analytical work of criminal analysts. Based on the behavioural profile of participants in crime activity, this paper is devoted to find the relationships between the actors. This paper is listing the basic principles required to build a real time intellectual system for crime analysis.

**2.3 Spatio-temporal crime prediction in rural:** This paper[3] proposes a spatio temporal predictive model which finds crime dense regions using k-means and DB scan clustering algorithms, extracts crime predictors using Seasonal Auto Regressive Moving Average model(SARIMA).The results and its implications are presented using appropriate graphs.

**2.4 An overview on Crime prediction methods:** In this paper[4], a detailed analysis of pros and cons of crime prediction methods are presented. The methods discussed in the paper are Support Vector machine (SVM), Fuzzy methods, Artificial Neural networks and multivariate time series algorithms for time dependant data. The author also suggested that hybrid methods may work better for crime prediction rather than filter methods.

## III. DATA UNDERSTANDING AND PREPROCESSING

**A. Data set:** The data set recorded the mere count for 30 different types of crimes that happened district wise for every state in India. This data set spans from the year 2001 to 2013. The data set is downloaded from kaggle and is already cleaned.

Further aggregations are done based on requirements. The dataset contains 9017 rows and 33 columns. The data set contains numerical value that represents the count/number of occurrences of that specific crime in a particular place and it is appropriate for descriptive analytics.

The values of the columns 'murder', 'Attempt to murder' and 'Culpable homicide not amounting to murder' are summed up in to a single column 'All murder' and those columns are dropped from data set, for the sake of clarity and simplicity. Similarly the columns 'Dacoity', 'Preparation and assembly for dacoity' and 'robbery' are dropped. Before dropping those, they are summed up in to a column 'All Robbery'. The columns such as 'Custodial rape', 'other rape', 'Kidnapping and abduction of women and girls', Kidnapping and abduction of others' and 'Importation of girls from foreign countries' are dropped from the table, since its values are meagre and not adding any significance to the analysis. The central tendency characteristics of data set are displayed as below snapshot.

| | RAPE | KIDNAPPING & ABDUCTION | BURGLARY | THEFT | RIOTS | CRIMINAL BREACH OF TRUST | CHEATING | COUNTERFIETING | ARSON |
|---|---|---|---|---|---|---|---|---|---|
| count | 9385.0 | 9385.0 | 9385.0 | 9385.0 | 9385.0 | 9385.0 | 9385.0 | 9385.0 | 9385.0 |
| mean | 29.1 | 45.0 | 130.2 | 412.7 | 90.3 | 21.7 | 93.2 | 3.1 | 13.1 |
| std | 32.3 | 63.3 | 195.7 | 786.9 | 138.8 | 37.8 | 191.5 | 8.4 | 29.5 |
| min | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25% | 8.0 | 10.0 | 28.0 | 86.0 | 10.0 | 3.0 | 12.0 | 0.0 | 2.0 |
| 50% | 20.0 | 25.0 | 80.0 | 208.0 | 44.0 | 11.0 | 37.0 | 1.0 | 8.0 |
| 75% | 40.0 | 56.0 | 165.0 | 434.0 | 115.0 | 25.0 | 99.0 | 3.0 | 18.0 |
| max | 568.0 | 923.0 | 3175.0 | 13195.0 | 3181.0 | 697.0 | 3155.0 | 217.0 | 2350.0 |

**Fig 1: Central Tendency of data set for some features**

### B. Bar Graphs

Bar graphs is one of the visual tools and graphical display that provides comparison of data among categories. The following bar graph clearly depicts that there is a steady increase in crime rate from the year 2001 to 2012.

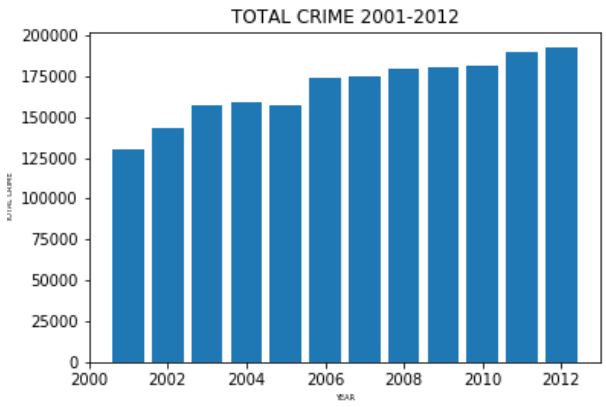



**Fig 2: Bar graph for total no of crimes vs. Year**

### C. Probability Density function:

Univariate analysis is quite possible using pdf. It is a frequency measure that gives the relative likelihood of random variable(X) in a given sample interval. In fact, Probability density function is a smoothed histogram of jagged points.

PDF F(X) = Frequency of X/ Length of the given interval

We have plotted here the frequency of 'All murder' on 'X' axis against length of Frequency on 'Y' axis. For instance, the probability of frequency of all murder at X= 200 is high and at X=250 is too low.

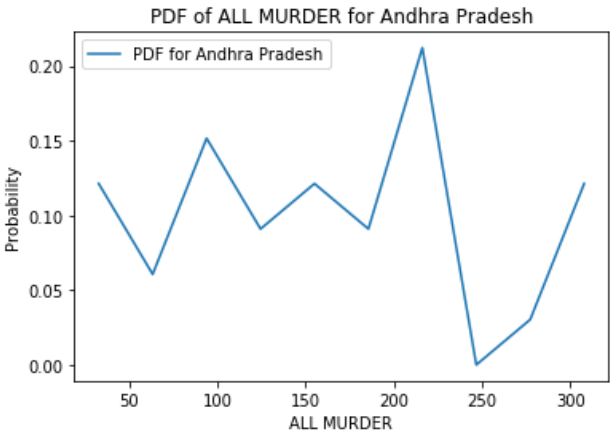



**Figure 3: PDF for feature "ALL MURDER'**

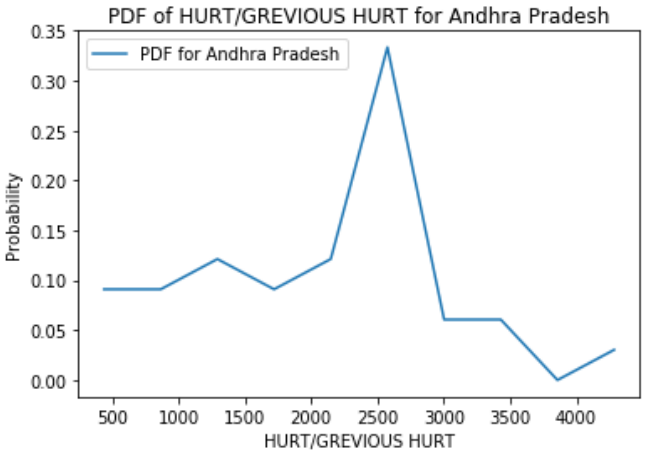



**Figure 4: PDF for feature "HURT/GRIEVOUS HURT" with a stark rise of probability for the value X = 2500.**

## IV. PROPOSED METHODOLOGY

### A. Unsupervised Learning

Unsupervised learning is one of the paradigms of machine learning where learning happens automatically without any Teacher. Learning happens from the unlabelled set. It is used to infer useful but hidden patterns from the unlabelled dataset. One of the most explored strategies of unsupervised learning is cluster analysis.

### B. Scatter Plot

**Scatter Plot**[13] is a two dimensional visualization of data that are plotted between two variables. Scatter plots depict the relationship between two variables. The frequency of crime type 'Rape' are scattered against the Total_ipc crime. Scatter plot shows the possible presence of clusters.

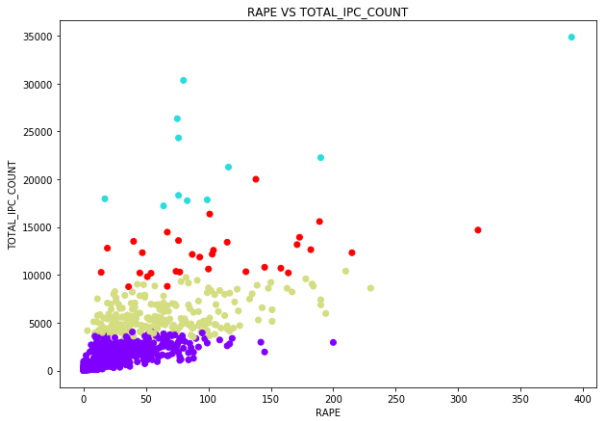



**Fig 6: Scatter plot for Rape Vs Total_IPC_Count**

But there are few data points which may be misunderstood as outliers except that, those are high count value for 'rape' in cities. Hence those points can't be omitted. From scatter plot, It is understood that building dendrograms to analyse the number of clusters possible is mandatory.

**Input:** Crime Data set CDS[m,n] with 'm' rows and 'n' columns
**Output:** Different Crime rated Zones $CZ_i = \{CZ_1, CZ_2, \ldots CZ_n\}$
**Algorithm:**
Begin

L1. Find dendrograms by applying hierarchical clustering method and identify the clusters.
L2. Apply elbow method on CDS to find 'n' clusters that are ideal and also confirms the number of clusters obtained from dendrograms
L3. Apply AGNES clustering algorithm on the dataset CDS, such that K = n to identify clusters c1,c2,….cn and label them using one hot encoding

L4. Label c1,c2,…cn as $CZ_1, CZ_2, \ldots CZ_n$.

L5. Find important features that cluster the data by building decision tree that uses Gini index

L6. Accuracy of decision tree algorithm is measured using log loss

Return $(n, CZ_1, CZ_2, \ldots CZ_n)$
End

Dendrogram is a tree like structure, which is a result of hierarchical clustering. Agglomerative nesting is one of hierarchical clustering algorithms. It begins with each data point in a cluster. At every step, two similar clusters are merged in to single cluster and is repeated till it becomes a single cluster containing all data points. This algorithm is run on the dataset to build dendrograms.

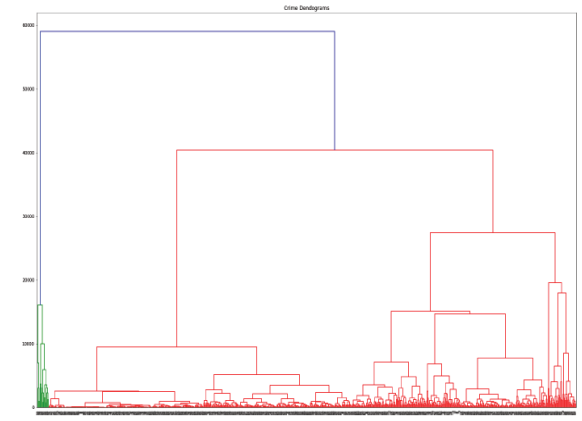



**Fig 7: Dendrogram for the dataset to find natural clusters**

In this dendrogram, the vertical axis displays dissimilarity index between clusters whereas horizontal axis displays the number of clusters. It is inferred from this dendrogram that it is ideal to find 3 clusters in the given data set at the dissimilarity height of 50000.

To support the evidence from dendrograms and to validate the ideal number of clusters, elbow plot is deployed. To build elbow plot, k-means clustering must be run on the data set for a specific range of k-values, say k = {1,2,…n}. K-means clustering [5] is used to cluster the data set and aims to partition the given dataset into 'k' partitions where each data belongs to one cluster. The data point is similar to the points of cluster, that it belongs to and is dissimilar to clusters that it does not belong to.

A line chart is plotted between 'k' and 'SSE' (Sum of Squared errors).The SSE can be calculated as the sum of squared distances from each point to its centroid. The plot looks like an arm and the point of inflection on the line chart is called elbow plot. Thus the point gives the ideal number of clusters on the data set. The elbow plot[15] given below shows the ideal number of clusters (K= 3 ) on the given data set.


**Fig 8: Elbow plot to find number of clusters**

## V. RESULTS AND DISCUSSION

We applied AGNES hierarchical clustering to find similar regions/cluster of districts in India based on the number of clusters inferred from elbow plot.
Each cluster represents the regions in India of similar crime rate. In hierarchical clustering, the distance between the points is calculated based on Euclidean distance measure as given.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)2}$$

The following word clouds are the clusters identified. The word cloud contains the name of the districts that belong to the same cluster.

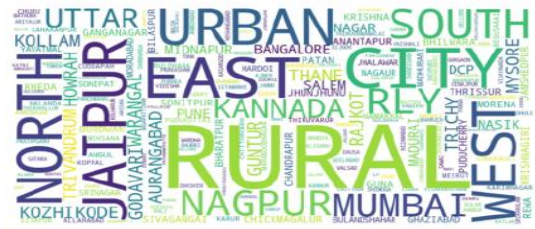
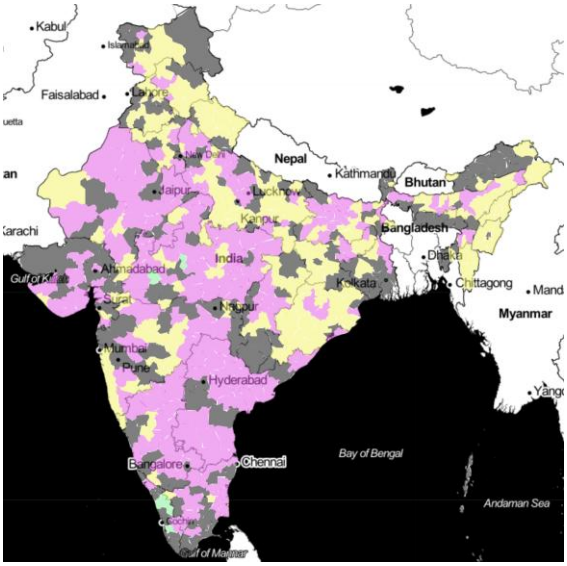



**Fig 9: Cluster 0**


**Fig 10: Cluster 1**


**Fig 11: Cluster 2**

A heat map of India is generated displaying three differently categorized crime zones of India using different colour codes.




## A. Finding feature importance

In order to find highly impacting features that are responsible for clustering data, a decision tree is built. It is a tree structure where in each internal node is an attribute or feature of the dataset and each leaf node corresponds to a label. The decision tree[12] is built using Gini index criterion[12]. This tree is built to determine which attribute among 'n' attributes is the 'Best split' at every level of decision tree. In simple, it determines the importance of feature at every level.

$$\text{Gini} = 1 - \sum_{i=1}^{c}(Pi)2 \quad \text{Where 'i' is a cluster}$$

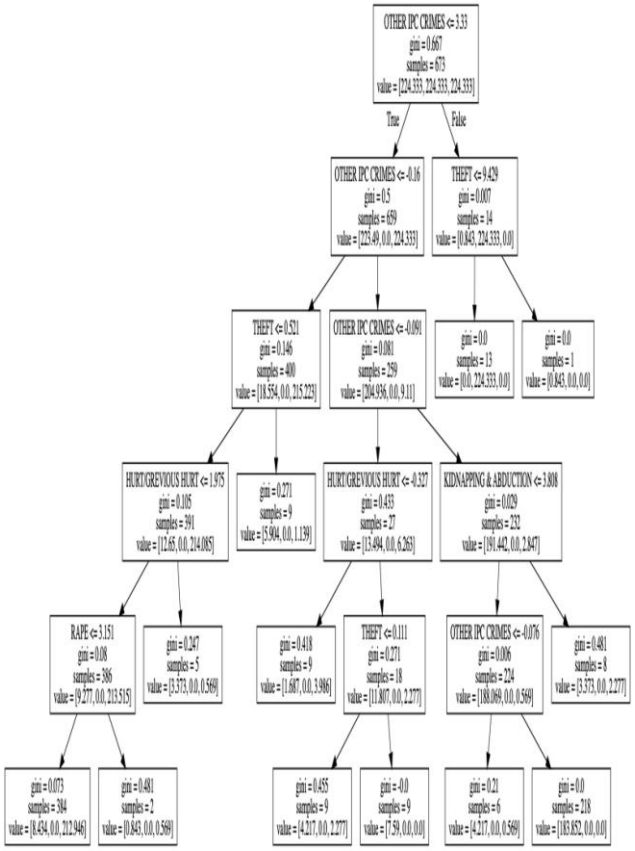



**Fig 13: Decision tree with high impact features**

From the decision tree, the below table shows the features in the decreasing order of significance that led to clustering.

**Table-I: Features in the decreasing order of importance**

| Sl.no | Features | Importance |
|---|---|---|
| 1 | Other IPC crimes | 0.949888 |
| 2 | Theft | 0.021974 |
| 3 | Hurt/Grievous hurt | 0.017865 |
| 4 | Kidnapping and abduction | 0.004182 |
| 5 | Cheating | 0.004003 |
| 6 | Rape | 0.002089 |

The accuracy of the built decision tree is measured using the metric logarithmic loss or log loss[15]. This metric is chosen for the reason that data set contained multiple class labels. It is a probability measure that tells how far the predicted label is varied from actual. This is an intuitive performance measure of a model. As the probability of prediction increases, log loss decreases and vice versa. Log loss is calculated as per the given formula and tabulated.

**Log loss = -1/n $\sum_{i=1}^{n}\sum_{j=1}^{c} Y_{ij} \log P_{ij}$ where 'i' is an instance and j is label**

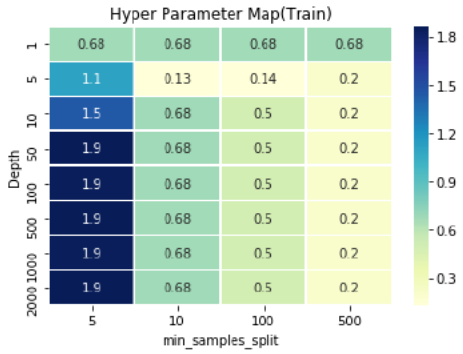



**Fig 14: Hyper parameter map on Training data set**

From the above results, it is inferred that an ideal decision tree can be constructed with depth 5 and at least for the minimum sample split of 10 at minimum log loss value of 0.13.

## VI. CONCLUSIONS

This paper is concluded to provide a systematic way of analysing the data set and to partition the entire districts of India in to three different zones based on crime level. There is a necessity to increase or withdraw the policing activity based on the crime rate at a particular location. The methodology proposed here is adding suggestions onto the rising or dropping the policing activity. It is also realised that policing and police strength that is successful of a particular village/ town/city of one cluster can be extended for other parts of the same cluster for better safety of public. It is determined that the other IPC crimes is one of critical features that clusters the districts of India.

The data set used for this paper is the best suited for descriptive analytics rather than predictive analytics. The features of the data set are independent of each of other. The data set contained mere count record of every type of crime. The drawbacks are the lack of dependant features that restrain us to find relation between the variables. The future work of this paper is to collect the deterrent or demographic variables such as literacy rate, socio economic status and schools per capita that influence crime activities to build an effective predictive model.

## REFERENCES

1. Manish Kumar, Athulya S, Mary Minu MB, Vidya Vinodini M D, Aiswaria Lakshmi K G, Anjana S, Manojkumar TK*, Forecasting of Annual Crime Rate in India:A case Study in : 2018 IEEE, pp.2087-2092
2. Dmytro Uzlov, Oleksii Vlasov, Volodymyr Strukov, Using Data Mining for Intelligence-Led Policing and Crime Analysis in 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology, pp.499-502
3. Charlie Catlett, Eugenio Cesario , Domenico Talia , Andrea Vinci , Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments: Pervasive and Mobile Computing 53 (2019) 62–74
4. Nurul Hazwani Mohd Shamsuddin1, Nor Azizah Ali, Razana Alwee, sFaculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia, "An Overview on Crime Prediction Methods", 2017.
5. Anant Joshi, A. Sai Sabitha, Tanupriya Choudhury, Amity University, Uttar Pradesh, HoD IT, Amity University, Uttar Pradesh, Assistant Professor, Amity University, Uttar Pradesh "Crime Analysis using K-means Clustering", 2017.
6. Vinit Kumar Gunjan1, Amit Kumar, Sharda Avdhanam, TATA Consultancy Services Ltd Hyderabad, Andhra Pradesh, India, "A Survey of Cyber Crime in India", 2013.
7. Shubham Agarwal, Lavish Yadav, and Manish K Thakur Department of Computer Science Engineering and Information Technology Jaypee Institute of Information Technology, Noida, India, "Crime Prediction based on Statistical Models", 2018.
8. Muhammad Umair, "An Overview of Crimes against Women and Children in Pakistan", Journal of Public Policy and Administration, 2018.
9. Hemraj Saini, Yerra Shankar Rao, T. C. Panda, International Journal of Engineering Research and Applications (IJERA), "Cyber-Crimes and their Impacts: A Review", 2012.
10. Rizqiya Windy Saputra, School of Electrical Engineering and Informatics Bandung Institute of Technology, sIndonesia, "A Survey of Cyber Crime in Indonesia", 2016.
11. Priyanka Das, Asit Kumar Das, Dept. of Computer Science and Technology Indian Institute of Engineering Science and Technology, Shibpur, Howrah, "A Two-stage Approach of Named-Entity Recognition for crime Analysis", 2017.
12. http:/dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/
13. https://chartio.com/learn/dashboards-and-charts/what-is-a-scatter-plot/
14. https://bl.ocks.org/rpgove/0060ff3b656618e9136b
15. https://datawookie.netlify.com/blog/2015/12/making-sense-of-logarithmic-loss/

## AUTHORS PROFILE

**J Vimala Devi** is a B.E graduate in Computer Science and Engineering from Bharathidasan University, M.Tech in Computer Science and Engineering from Satyabama University, Chennai and currently pursuing my Ph.D in Visveswrays Technology University. She has 13 years of Experience in Teaching. She is a life member of CSI and ISTE. She published a paper titled "Fraud detection in Credit Card Transactions by using Classification Algorithms " in IEEE conference. She is doing a lot of projects on title "Descriptive analytics". She is currently working as Associate Professor, Department of CSE in Cambridge Institute of Technology, Bangalore.

**Dr Kavitha K S** completed BE computer science and Engineering from SIT Tumkur, MTECH from BMSCE Bangalore and PhD from Anna University, Chennai. She is having 22 years of experience in engineering colleges and has worked in various capacities. Around 40 research papers under her credit. Currently she is supervising 6 research scholars. Her areas of interest are Data mining, Machine Learning, Algorithms and Programming etc. Currently she is serving as Professor in the Department of CSE, Global Academy of Technology, Bangalore. She is a life time member of professional bodies such as CSI and IEEE.