K-Means and Hierarchical based Clustering in Suicide Analysis

R Sujatha, S. Sree Dharinya, E P Ephzibah, R Kiruba Thangam

Abstract: Machine learning is the intriguing area of research that spreads across all domains helping in providing quality decisions. Demographic have more influence in social happenings along with various personal and social factors. Suicide analysis is one such issue to be handled with great concern that will provide precautionary based on situations. Suicide prediction can be carried on using data mining that can be used to predict the suicide earlier so that it can be prevented. Suicide is an action resulting in death performed by themselves. Common factors that influence the rate of suicides are cause, method of suicide, year, gender, educational qualification, social status. For this clustering technique in datamining that falls under unsupervised provides great platform. Silhouette score is used for mapping the number of cluster to get the good clustering. Various plots like box plot, scatter plot and so on helps to provide greater insight. Based on analysis the required remedial could be arrived.

Index Terms: Clustering, Suicide, Gender, Age, Education,

I. INTRODUCTION

Machine learning is the boon to various field in making strategy decision. Based on the type of data and size of data in the recent years several algorithms have been populated to serve the purpose of decision makers. The volume and variety of data made it potential research area. Supervised, unsupervised and reinforcement learning are the broad categories in machine learning that functions based on the input, output or target variables. The other milestone in the variables are dependent or independent variable along with the level of dependency among variables if it falling in the criteria of dependent variables. Clustering is interesting area in the unsupervised learning methodology [1].

In the modern era people living in world face different problems in midst of competitive space. The quantum of withstanding differs from time to time, place to place, gender to gender, profession to profession, and level of education to level of education and more importantly social status. This depression and inequalities pave way to take wrong decision in shorter span of time hastily. The wrong decision is ending the life without thinking about facing challenges in optimistic manner. That decision is suicide. The promptness in ending the issue by self-killing. The trend of suicide across the globe as increased exponentially and recent research mentioned by world health organization states that every year approximately 1 million people die. It's at the alarming rate of 1 death per 40 seconds [2]. Both existing data along with clinician only the suicide threat can be combatted [3]. Based on the available data strong statistical analysis along with datamining techniques definitely provides better platform to curtail the issues that could arise of adverse decisions of general public [4][5]. In the context of this seen a dataset that comprised of suicide incidents across the India in all the states and union territories. The dataset is quiet hefty with long duration from 2001 to 2012 [6]. The structure of dataset is as follows [7].

For each state and union territory for the years mentioned with causes for suicide for both genders in different age group with total number of suicides are illustrated as follows
Death of Dear Person
Divorce
Not having Children (Barrenness/Impotency
Drug Abuse/Addiction
Other Causes (Please Specify)
Paralysis
Poverty
Suspected/Illicit Relation
Cancellation/Non-Settlement of Marriage
Illegitimate Pregnancy
Causes not known
Failure in Examination
Other Prolonged Illness
Fall in Social Reputation
Unemployment
Love Affairs
Illness (Aids/STD)
Insanity/Mental Illness
Professional/Career Problem
Dowry Dispute
Ideological Causes/Hero Worshiping
Bankruptcy or Sudden change in Economic
Property Dispute
Cancer
Family Problems
Physical Abuse (Rape/Incest Etc.)
Similarly in place of causes the education status for each state and union territory for the years mentioned for both
genders in different age group with total number of suicides are illustrated as follows
Graduate
Hr. Secondary/Intermediate/Pre-University
Post Graduate and Above
Diploma
Middle

No Education
Primary
Matriculate/Secondary
Similarly with means of adopted for committing suicide for each state and union territory for the years mentioned for both genders in different age group with total number of suicides are illustrated as follows
By Other means (please specify)
By touching electric wires
By Jumping from (Building)
By Drowning
By Machine
By Consuming Insecticides
By Consuming Other Poison
By Self Infliction of injury
By Overdose of sleeping pills
By Fire/Self Immolation
By Over Alcoholism
By Jumping from (Other sites)
By Jumping off Moving Vehicles/Trains
By coming under running vehicles/trains
By Fire-Arms
By Hanging
The other important factor is the professional status
Others (Please Specify)
Professional Activity
Others (Please Specify)
Farming/Agriculture ActivityService (Government)
Public Sector Undertaking
Service (Private)
Unemployed
House Wife
Student
Self-employed (Business activity)
Retired Person
The very sensitive and emotional attribute of every individual based on social status
Married
Separated
Widowed/Widower
Never Married
Divorcee
The dataset is processed and as sample concentrated on the Tamil Nadu state. Applied various statistical and clustering techniques to retrieve interesting and informative decisions to prevent suicides.

II. PROPOSED METHODOLOGY

A. AGE Based Analysis

The death caused by suicide is not uniform in all age groups. It differ greatly in different age groups. The death ranges from 3205 in 0-14 age group to 56599 in 30-44 age group. The study of graph produced by data based on age gives us the figures about death caused by suicide in different age groups. From the graph we can conclude that most of the suicide are in age groups of 30-44 while the age group of 0-14 is least. The second highest is 15-29 age group. The age group 60+ that is retired people suicide values are much less than top values. These are directly associated with the stresses of people of age groups 15-59. In the first age group most people are school students who don't have much stress. The second third and fourth age groups mostly have people of working class who have many sorts of stresses like family, job, economic, social. The fifth age group consists retired people who will have less stress as compared to other groups. So the Suicide rates are profoundly impacted by the age. Figure 1 shows the number of suicides among different age group [8].
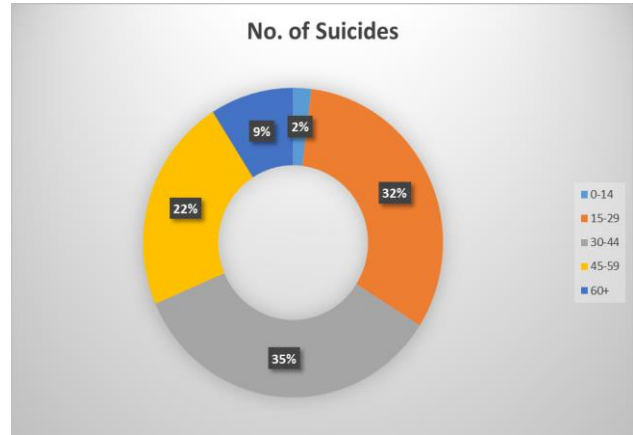

Fig.1. No. of Suicides among Age Group

B. Education Based Analysis

Education is powerful weapon to change the world. It can influence people greatly. So the education can influence people decision making on suicide. The study shows that lacking of education can lead to more suicides. The suicide counts are least in high education qualifications and most in no education. This suggests that educated people can take productive better decisions. So educating the people can reduce the rates of Suicide. Figure 2 provides insight about number of suicides based on educational background.
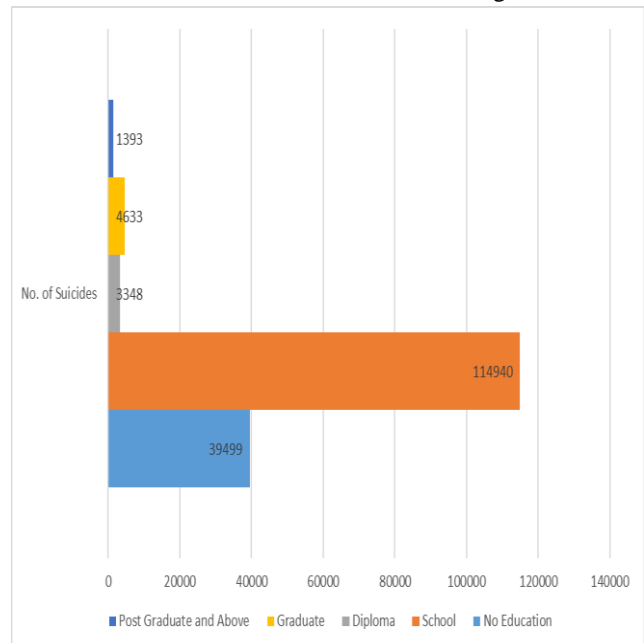

Fig.2. No. of Suicided Based on Education

406

C. Social Status Based Analysis

The social status plays key role in any society. The suicide numbers of different social groups vary greatly. Even in the same social group the rate of suicide in male and female differ by large number. In divorcee the number is suicide is relatively equal in both male and female. In married the suicide rate in male is approximately double the suicide rate in female. In never married suicide rate is higher in male as compared to female. In separated also male suicide rate is double the female suicide rate. In widowed the suicide rate in female is many times the suicide rate in male. The highest rate of suicide is in category of married as they create more responsibilities. The study in suicide in social perspective suggests that widowed female require special care. Figure 3 shows the social status with gender information along with number of suicides.
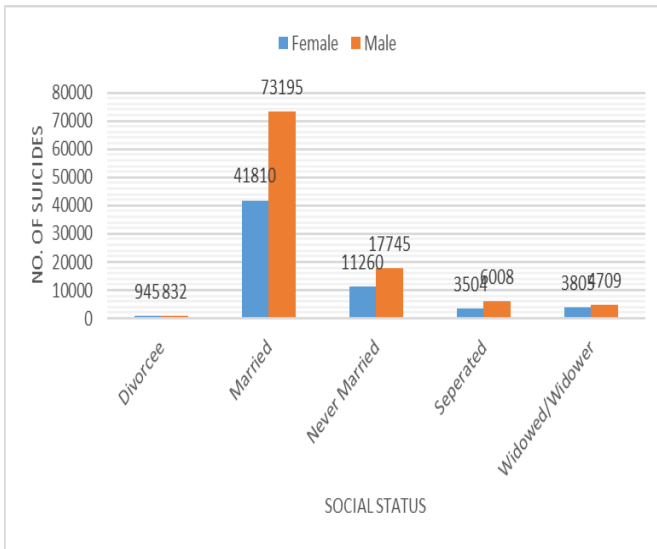

Fig.3. No. of Suicided Based on Social Status

III. PROPOSED WORK

With the help of ranking method with information gain features are selected. K-Means clustering is the very old technique used to build the clusters [9]. The objective of grouping is putting together like valued things. So that grouping and taking decision will be profitable [10]. The complete work flow of the proposed work is illustrated in figure 4.
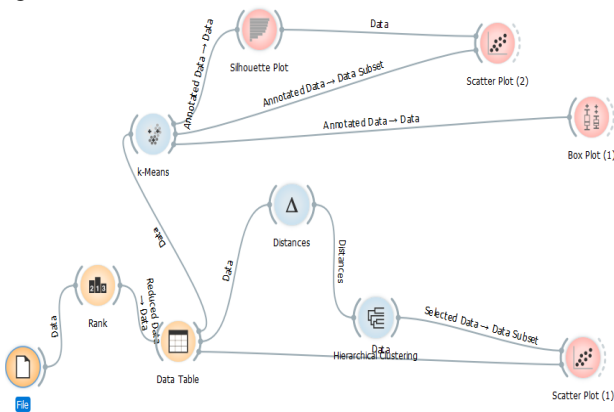

Fig.4. Proposed Work Flow

A. Silhouette Score

Silhouette is a method for interpreting and validating consistency of clustering [11]. The score ranges between -1 to +1. High value indicates the object is well matched and in turn good clustering is achieved. It can be computed with any distance metric such as manhattan and Euclidean. Silhouette score for varied clusters have been computed and for cluster 3 it's the best and based on the scatter plot is plotted and from K-means box plot for age, gender and we can make for all. Figure 5 shows silhouette plot. Figure 6 and 7 shows box plot based on gender and age. Figure 8 shows the scatter plot for K-means with silhouette score based on profession, age and number of suicides.
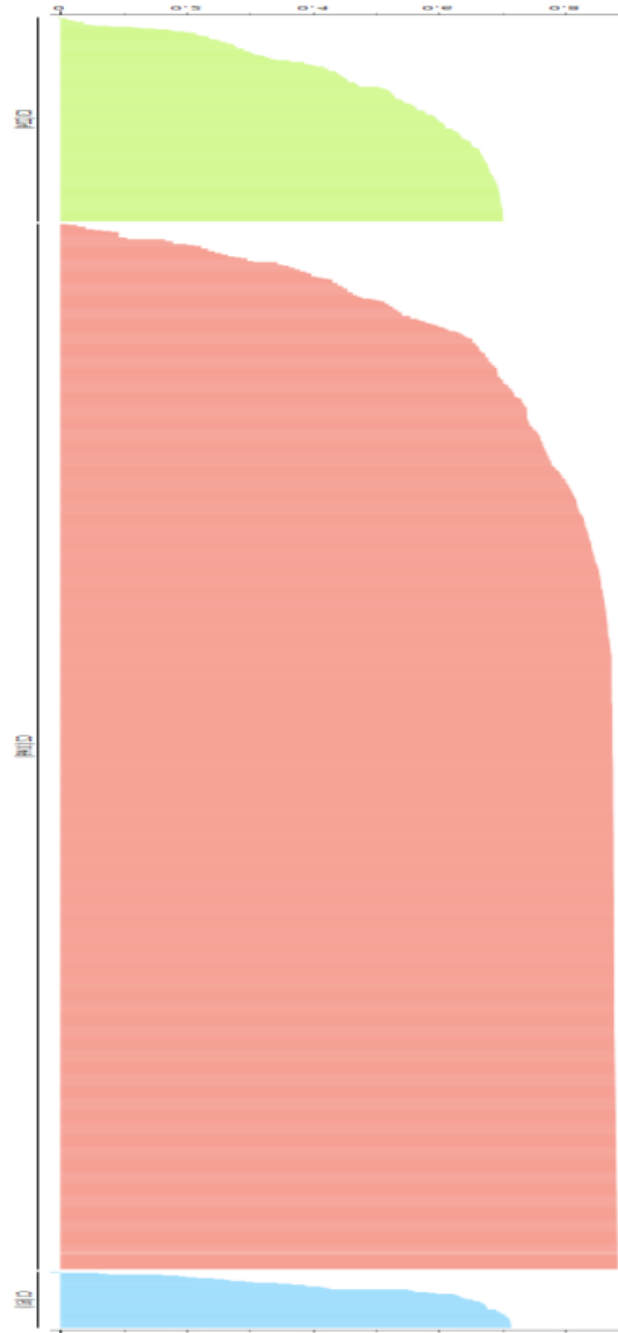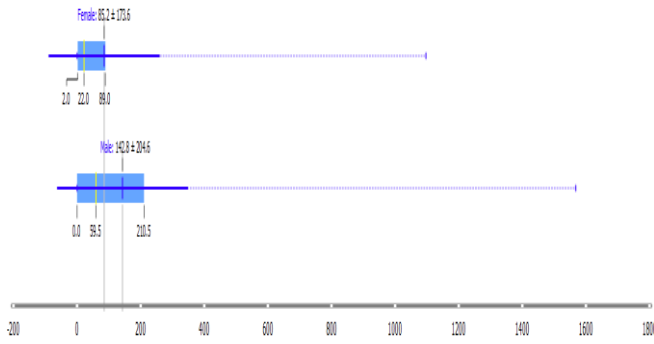

Fig.5. Silhouette Plot
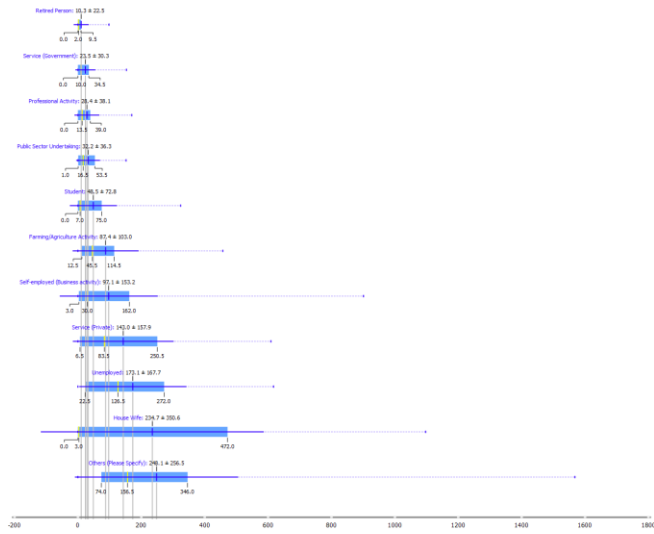
Fig.6. Box Plot - Gender
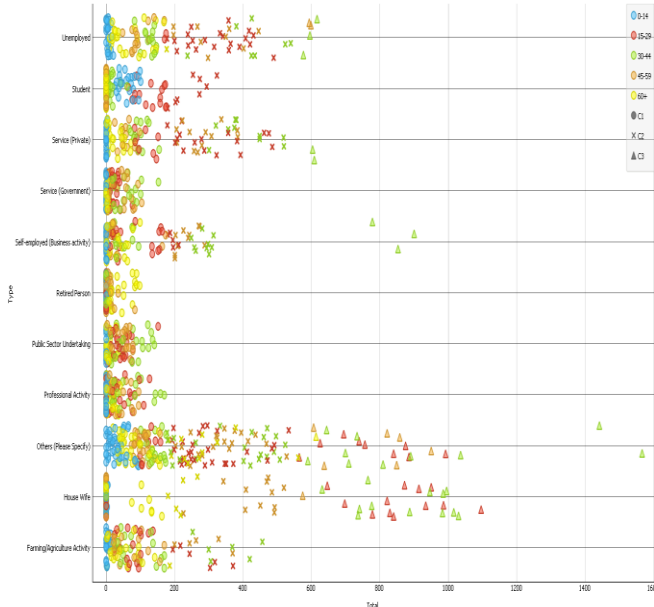

Fig.7. Box Plot – Age

Fig.8. Scatter Plot

B. Hierarchial Clustering

Most suicide victims we the method of self-killing. Their behaviours are mostly self-injurious in pre suicide period. So the activities of people can be collected and clustered to predict suicide [12]. Clustering method that is used to build cluster tree also popularly referred as dendrogram is called as hierarchical clustering. Similarly scatter plot drawn for the same.

IV. RESULTS

On the educational viewpoint, Illiteracy is directly proportional to suicidal rate and effective measures required to empower all with knowledge is required. In the fast growing world the tolerance level of people effects their personal life and it's evident that in married social status the number of suicides are higher. Working people on the middle age group more prone for self-killing. Causes and method adopted varied based on the situation.

V.CONCLUSION

The mental illness cause large number of deaths by suicide. The people suffering from psychiatric disorders don't have self-control they expose too much of all feelings. So when they are frustrated they hurt others or themselves. The sudden events are referred to precipitating events that makes the worse condition [13]. Development of suitable tools will greatly reduce the riskiness based on the historical data analysis. The application of artificial intelligence can help to reduce suicide rates because it can predict the Suicides earlier. The artificial intelligence can be a good friend for people in times of hardship to glide over hard time without taking wrong decisions like suicide [14]. The text message processing can be used to earlier prediction of suicide. Today people use text messaging platforms more to express themselves to others. So the text messages can help to identify behaviour which can lead to suicide in near future. The actions to prevent suicide are need of an hour [15]. So development of risk assessment tools can play highly positive role in reducing the suicide level. These tools monitor every moment of people [16].

REFERENCES

1. Linthicum, Kathryn P., Katherine Musacchio Schafer, and Jessica D. Ribeiro. "Machine learning in suicide science: Applications and ethics." Behavioral sciences & the law(2019).
2. https://www.befrienders.org/suicide-statistics
3. Beck, Aaron T., Dan J. Lettieri, and Harvey LP Resnik, eds. The prediction of suicide. Bowie, MD: Charles Press Publishers, 1974.
4. Larose, Daniel T. Data mining methods & models. John Wiley & Sons, 2006.
5. Mandge, Omprakash L., and Bandra Reclamation. "A data mining tool for prediction of suicides among students." In Proceedings of National Conference on New Horizons in IT-NCNHIT, p. 178. 2013.
6. https://www.kaggle.com/rajanand/suicides-in-india
7. Anil, R. A. N. E., and Abhijit Nadkarni. "Suicide in India: a systematic review." Shanghai archives of psychiatry 26, no. 2 (2014): 69.
8. Beautrais, Annette L. "Risk factors for suicide and attempted suicide among young people." Australian & New Zealand Journal of Psychiatry 34, no. 3 (2000): 420-436.
9. Capó, Marco, Aritz Pérez, and Jose A. Lozano. "An efficient K-means clustering algorithm for massive data." arXiv preprint arXiv:1801.02949 (2018).
10. Choo, Carol, Joachim Diederich, Insu Song, and Roger Ho. "Cluster analysis reveals risk factors for repeated suicide attempts in a multi-ethnic Asian population." Asian journal of psychiatry 8 (2014): 38-42.

11. https://en.wikipedia.org/wiki/Silhouette_(clustering) Accessed 20 February 2019.
12. Bani, Marco, Gabriele Travagin, Michele Monticelli, Manuela Valsecchi, Emanuele Truisi, Federico Zorzi, Mariagrazia Strepparava, Massimo Clerici, Umberto Mazza, and Giorgio Rezzonico. "Pattern of self-injurious behavior and suicide attempts in Italian custodial inmates: A cluster analysis approach." International Journal of Law and Psychiatry 64 (2019): 1-7.
13. Maltsberger, John T., Herbert Hendin, Ann Pollinger Haas, and Alan Lipschitz. "Determination of precipitating events in the suicide of psychiatric patients." Suicide and Life-Threatening Behavior 33, no. 2 (2003): 111-119.
14. Marks, Mason. "Artificial Intelligence Based Suicide Prediction." Yale Journal of Health Policy, Law, and Ethics, Forthcoming (2019).
15. Velupillai, Sumithra, Gergö Hadlaczky, Enrique Baca-Garcia, Genevieve M. Gorrell, Nomi Werbeloff, Dong Nguyen, Rashmi Patel et al. "Risk Assessment Tools and Data-driven Approaches for Predicting and Preventing Suicidal Behaviour." Frontiers in Psychiatry 10 (2019): 36.
16. Glenn, Jeffrey J., Alicia Nobles, Laura Barnes, and Bethany Teachman. "Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally-Sensitive Markers of Suicide Risk." (2019).

AUTHORS PROFILE

Dr. R. Sujatha is serving as an Associate Professor in Vellore Institute of Technology. Possess 15 years of teaching experience. She has completed her B.E. degree in computer science from Madras University, in 2001, the M.E. degree in computer science from Anna University in 2009 with university rank and finished her Ph.D. in computer science in 2017 in Vellore Institute of Technology, Vellore. Wrote a book on software project management for engineering students. Published over 30 journals and 10 conferences. Actively participate in administrative tasks in university. Acted as an invited speaker and coordinated various program. Quiet interested in listening to songs, gardening, trying different dishes and hand works. Her area of research interest includes Data mining, Management of Information systems, Software Engineering, Big Data, and Analytics.

Dr. S. Sree Dharinya is working as an Associate Professor, in Vellore Institute of Technology, Vellore. She earned her PhD from VIT University in Information Technology. She holds a B.E in Electrical and Electronics Engineering and M.S in Software Systems .She has 16 years of teaching experience and her research interest includes software engineering, machine learning, sentiment analysis, and big data analytics.

Dr. E. P. Ephzibah received her Ph.D (2016) from mother Teresa Women's university, kodaikanal, India. She is currently working as Associate professor in the School of information technology and engineering, VIT, Vellore. She has over 15 years of academic and research experience. Her research area includes artificial intelligence, soft computing techniques, block chain technology, classification, clustering and prediction.

Prof. R. Kiruba Thangam ,Assistant Professor[Sr.] at Department of Software and Systems Engineering ,School of Information Technology and Engineering, VIT University .She has 14 years of teaching experience. Her research interest includes Image Processing, Data mining and Software engineering.