ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.722-728, October 2020

Big Data Analytics on Matrimonial Data Set

Anurag Sinha[1], Arinjay Biswas[2], Tushar Raj[3], Aditya Misra[4]
[1,2,4] Department of Computer science and IT, Student, Amity University Jharkhand
Ranchi, Jharkhand, India
[3]Department of Mathematics, Student, Birla Institute of Technology, Mesra 835215 Ranchi
Email: anuragsinha257@gmail.com[1],biswas87280@gmail.com[2],tushar4460@gmail.com[3],
misra_aditya098@hotmail.com[4]

Abstract-In India online matrimonial websites are widely used among the customers and are the trusted and reputed website of the Indian globally. The number of people getting registered in the matrimonial sites it creates a lot of data sets, a lot of unstructured data sets which is floating around the internet. there comes the concept of big data analysis, big data analysis is the analysis of the data science and machine learning algorithms through which we can extract and mine the knowledgeable and informative data sets which can be used for predicting the future instances. In this paper we are analyzing the big data set from the machine learning data repository, to be driven analyze the different it is it's using implementing a different machine learning algorithm and regal the informative information. we will be implementing various data mining operations such as opinion mining and sentiment analysis data visualisation and machine learning technique, marriage information have always been an integral and confidential knowledgeable in any civilized and well-educated society. some online agency rate on these matrimonial websites by creating a profile of an individual to get in touch with the other individual for their life purposes. in this paper, we will be analysing the total number of sexual harassment and domestic violence cases that have been registered in India. since 2001 marriage culture the dowry is one of the conventional and ever emergence kind of sickness. which cannot be removed ever. using data visualisation technique by a language we will show how many cases have been registered in India in sexual harassment and dowry. we will also implement the several operations such as horoscope preferences income and statuses of the individual as per the data collected by the matrimonial data repository.

Keywords – Big Data, Data Visualization, Machine Learning, K-Mean Clustering, Density-Based Clustering

I. INTRODUCTION

India is a secular democratic country consists of 28 states and 8 union territories, for a total of 36 entities. In the world, India has a vast diversity in its culture. India has a population of 1.3 Billion people which is the second-highest in the world. India has 23 official languages and remarkable religious differences of around 20 religions including Hinduism, Buddhism, Jainism, Islam and Sikhism. In India, there is a vast diversity in socioeconomic status about educational fulfilment, social power, gender inequality, urbanity, caste, etc. A daily wage labourer to a billionaire businessman, tribal illiterates to high-class intellectuals, slum dwellers to NRI and common people has received equal attention towards the formation of diverse groups of the nation. Since the last decade, India shows great participation in the new era of digitalization due to which we can see the huge number of matrimonial sites in India. Nowadays, matrimony sites help in finding a better match by filtering the search by racial, ethnic, linguistic and cultural diversity. Before matrimony sites, arranged marriages happened through the matrimonial column in newspapers or through relatives, priest and mutual friends. Nowadays, people are much busier than ever and they don't have time to find a suitable life partner for themselves through a traditional approach. Matrimonial sites seem to be the only way to find a better life partner without relying on relatives and marriage bureaus.

II. LITERATURE SURVEY

Arranged marriages in the past happened to be trough matrimonial columns in the newspaper or through the acquaintance of relatives or friends or priests. But then, the people working in India or abroad didn't have time to look for their match the traditional way. They wanted to blend old & new sensibilities to find the perfect match of their parent's choice.[1,2,3,4]

Thus, by time, the Matrimonial system came into the picture. An alternative to strike a comparison between ancient social tradition & the contemporary attribute by cutting the intermediary of arrange marriage. Through this matrimonial system, it has become uncomplicated for those who have never met or known each other, or are culturally different & live diagonally across the globe, become life partners.[4,5]

This literature survey studies the performance of classification algorithms based on the bride or groom data of the matrimonial system. The working process for each algorithm is analyzed with the accuracy of the classification algorithm. it also studies various data mining techniques applied in finding a suitable match for the respective bride or groom. In 2007 Aman and Suruchi have experimented in the WEKA environment by using four algorithms namely ID3, C 4.5, simple CART, and alternating decision tree on the student's dataset, and later these were compared in terms of classification accuracy. according to their simulation results, the C4.5 classifier outperforms the ID3, CART & AD Tree in terms of classification accuracy.

Abbott, D., Matkovsky, P. & Elder, J. (1998) [6,7] in 2007 presented analysis on accurate prediction of academic performance of undergraduate and postgraduate students of two very different academic institutes: Can Tho University (CTU), a large National University in Vietnam and Asian Institute of Technology (AIT), International postgraduate institute in Thailand. they have used different data mining tools to find the classification accuracy from Bayesian Networks and Decision tree. they have achieved the best prediction accuracy which is used to find the performance of students. the result of this study is very useful in finding the best-performing students to award with scholarships. The result of this research indicates the decision tree was consistently 3-12% more accurate than the Bayesian Network.

Sukonthip & Anorgnart in 2011 presented their study using data mining techniques to identify the bad behaviour of students in vocational education, classified by algorithms such as Navie Byes Classifier Bayesian Network, C4.5, and Ripper. Then it measures the performance of classification algorithms using 10 folds cross-validation. it is showed that the C4.5 algorithm is not appropriate for all data types, & the Bayesian Belief Network Algorithm that yields an accuracy of 82.4 per cent. Shilpa Dharkar, Anand Rajawat in 2012, proposed a recommendation system which is based on web data mining which is the application of data mining technique helping us to determine the pattern from the web. In terms of accuracy and time, performance analysis of learning algorithm ID3 and C4.5 apply it on healthy diet application.[8,9,10]

T. Miranda Lakshmi, A.Martin, R.Mumtaz Begum, Dr V Prasad Anand Venkatesan[11,12] In 2013 study about the performance of classification algorithms based on student data the working process for each algorithm is analyzed with the accuracy of classification algorithms. also studies various data mining techniques applied in finding the student academic performance using ID3, C4.5, and CART decision tree. In 2013, few people, Anuja Priyam, Abhijeet, Rahul Gupta, Anuj Rathee, and Saurav ShrivastavaProposed decision tree algorithm former applied on the data and the results are compared of all algorithms and evaluation is done by already existing datasets. In 2011, Dr K. Usha Raniand & D.Lavanya, presented a paper that was based on the performance of decision tree induction classifiers on various medical data sets, in terms of accuracy and time complexity were analyzed.

III. RELATED WORK

One category of related work is social science work. In this, we refer to the statistical prediction process and conclusion of the works as us, for a subjective method of feature selection. These works were based on social science, and that only does simple regression and tests the statistical significance. This included overfitting problems. In our case, we combine knowledge from these research works and then train models through machine learning algorithms[13].

The expert in marriage outcome research, Gottman, was able to construct a short term diverse prediction model for the couples who are 93% accurate. His model primarily uses features about marathon interaction and emotions. The negative effect during the marital conflict most predicted breakups. Are model deposit in that the feature set reuse primary consists of the demographic data. Besides, we predict the outcome of relationships in general, not only married couples. Finally, extensive cross-validation on the data set was done. the second category of related work considered the clinical machine learning problem because of the similar issues such as unbalanced data set Nokia boundary a large number of features and factor analysis issues etc. Thus, we do this work on feature selection and evaluation methods[14].

IV. METHODOLOGY

Data analysis

India has the old tradition of arranged marriages. in which the parent speak the most eligible candidate from another house and with the evidence that he is a good earner in his having the good income and based on the educational and societal qualification and status two range or unknown individuals get married so we can see that finding a partner is a significant decision for the long life perception. that is why they are several internet services and sites started a business of providing a platform in which the individuals can save their bio-data in the database in the form of their data educational data and statistical data on which the individual can access each one status and profile for being married. We have taken the Shadi.com data set to aggregate data that would provide the insights into above-mentioned social logical bijous. Honest survey it is found that that on this website the total number of males is 76.42 per cent and females numbering are having 23.5 8% shaadi.com is one of the leading Matrimonial website containing approximately 325000 profile Targets the educated urban and middle-class family for the matrimonial site which refers to India's world's largest metropolitan cities in India. Data analysis is done on the AI Python notebooks and Pandas data analysis library and NumPy frameworks. I am putting a screenshot of one data set of CSV file that how it can help in opinion mining and sentiment analysis[15].

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.722-728, October 2020

```
about1,about2,about_family,age2,birth1,birth2,birth3,birth4,body1,body2,brothers,caste1,caste2,city2,college,complexion1,complexion2,created_for,degree,
details,drinking1,drinking2,eating1,eating2,education2,employed1,employed2,family_type1,family_type2,father,fields,gotra,have_children,height1,height2,h
oroscope,income1,income2,last_login,living,manglik1,manglik2,marital_status1,marital_status2,mother,mother1,mother2,nakshatra,occupation,profileID,raasi
,religion1,religion2,sisters,smoking1,smoking2,special1,special2,status1,status2,subcaste,url,values1,values2,weight1,weight2

"Simple, loving, caring very emotional n sensitive by nature, respect feelings and emotions and expects the same. Currently working as a Quality
Associate at AMAZON.",,,25 to 30,29 May 1988,,,,Slim,,Kurmi Kshatriya,Kurmi Kshatriya,,,Wheatish ,,Daughter,MBA/PGDM,"25 yrs, 5ft 4in,<br /> Kurmi
Kshatriya, Hindu, Hindi ,<br /> MBA/PGDM, Others, <br /> Bangalore, India",No,,Vegetarian With Eggs,,,MNC,,,,HR/ Industrial Relations,,No,5ft 4in,5ft
4in to 6ft 1in,,INR 2 lakhs to 3 lakhs,, 08 Nov 2013,,No,No,Never Married,Never
Married,,Hindi,Hindi,,Others,SMF1760790,,Hindu,Hindu,,No,,None,,,Singh,,,,,
```

FIGURE 1: DATASET EXAMPLE

Text Analysis- Predicting Horoscope Preference

Approximately 140000 users having the texture section in their profiles and these approximate fields are analyzed by using the machine learning algorithm which is *naves bayes* to model. This method is used because it determines the text classification and the predictions which could be made very precisely. This is the kind of a classifier which follows the simple probabilistic classification based on the best theorem with Independence assumptions. A lot of redundant words are removed and represented in each profile is created using this package. As shown in the given figures visualization moving to a big city for getting a good education does not mean that the reduction of the trust and believe and forgive forgetting the value and custom and Hindu Vedic system. The view of horoscope matching is kind of determining the compatibility for the marriage. Words more likely to predict the Hindu Vedic horoscope based on the birth date and time are shown in this particular compatibility visualization[16].

Analysis - when referencing the horoscope matching is bifurcated into two different locations first examine the location under the visualization. Is it is determined that the horoscope matching is based on the users having the big cities correspondents as other from the user having from the small town and villages. the second clouds shows the few words which is education and career it mentions from the smaller cities and towns, it has few green adjectives which shows that the person having the *jovial cultured and well civilize.* kids to be because we uses in the first cloud our belongings from the big cities and awesome to have the better career is better education and better selective preferences in their criteria of the partners elections and thereby second horoscope matching is the additional creation as an optional background. User having in second cloud is bit less selective words which are highlighted in yellow like *deaf dumb condition and treatment* shows that that they perceive themselves as the lower value in the marriage market and therefore their belongings and the specific characteristics about the requirement. we can say that we are implementing the opinion mining and sentiment analysis of the data that we have on the matrimonial sites to gain be preferences of the horoscope based on the locality and their preferences and their characteristics and the words with show their own characteristics and which depicts their own ethics and values.

Data collection –

We have taken data of total no of cases registered in dowry from govt. data repository of dowry cases in India an collected matrimonial data from MACHINE LEARNING DATA REPOSTORY OF HARDVARD UNIVERSITY[17].

FIGURE 2: VISUALIZATION OF DATA

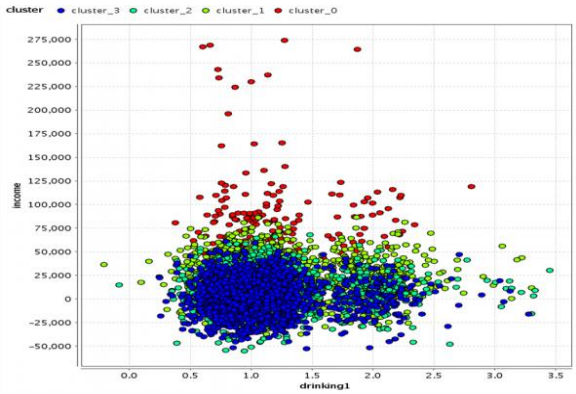
Figure 3: BAG OF WORDS SHOWING HOROSCOPE PREFERENCS

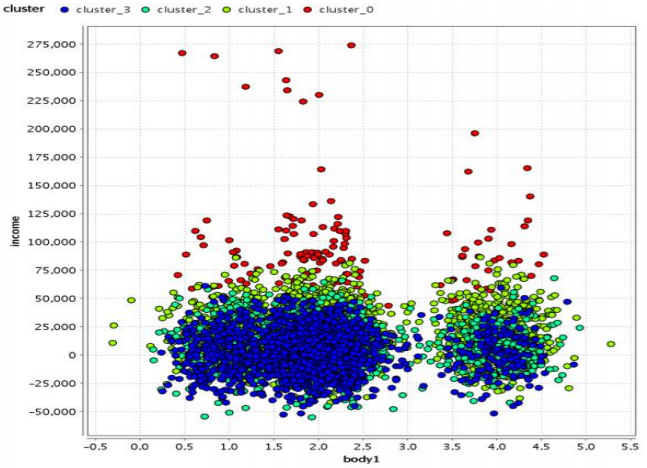
ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.722-728, October 2020

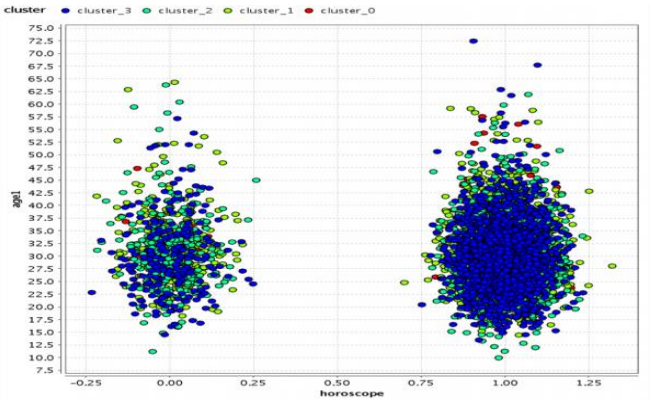
K-Means Clustering

I t seems that the marriage marketing which treats people as if they are on a leather and even the better attributes on the most valued then don't the votes attributes at the bottom. People have are having the most flexible Ness are having the most valued in the marriage marketing seems to be very selective out of the crowd. This particular analysis we have shown or table below in which we have shown the highest income highest education, lowest income, lowest education, uses seems to decide the affects the horoscope matching more than the other. k- Mean clustering is performed on the random subset of the data which is the available time the computational power using the Rapid Miner predictive analytics package. in the following clusters the mostly male are hired having the highest income and members are also the taller and more educated and tend to the more flexible to their profiles. Cluster 3 is having the comparatively few mail which having younger and lowest average income and are shorter and having the pitiless in this kind of profiles are in cluster 3. enclosed to the members tend to be tend to live in the joint family and regard the most important to the caste system and having the same links drinks often less or more closer to the traditional and conservative mind-set are kept in the cluster 2.

Visualizing the Clusters







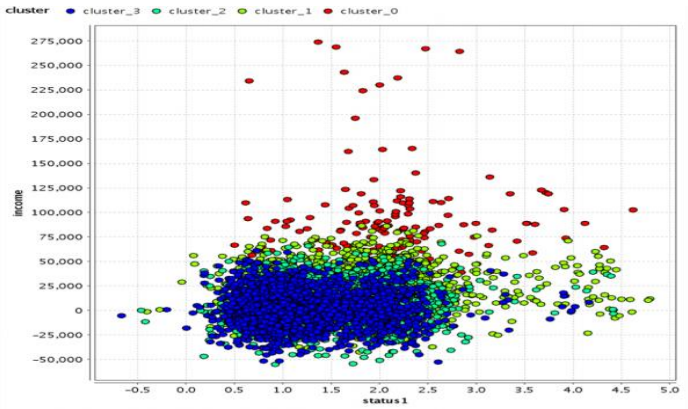


Figure 3 - Status: Middle Class: 1, Upper Middle Class: 2, High Class: 3, Rich/Affluent: and Horoscope: Don't want horoscope matching: 0, Want horoscope matching:1

Figure 4 - Living: Living alone: 0, Living with parents: 1, Mother: Armed Forces: 1, Business/Entrepreneur: 2, Civil Services: 3, Housewife:4, Passed:5, Retired:6, Service - Govt/ PSU:7, Service - Private: 8, Teacher:9, Not Employed:10, Body1: Doesn't Matter:0, Slim: 1, Average: 2, Heavy: 3, Athletic: 4

Density vase clustering -DbScan is a density-based clustering algorithm that finds a number of clusters starting from the estimated density distribution of corresponding nodes Here we use the Rapid Miner predictive analytics tool to run DbScan on a table consisting of body1 (body type, where Doesn't Matter:0, Slim: 1, Average: 2, Heavy: 3, Athletic: 4) versus BMI. Three clusters are generated: Cluster 1: Average, Cluster 2: Slim, Cluster 3: Athletic, Cluster 0: Unclustered ("heavy" comes in this category). Here is the number of members per cluster when the clustering is run on a random sample size of 5865 people):

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.722-728, October 2020

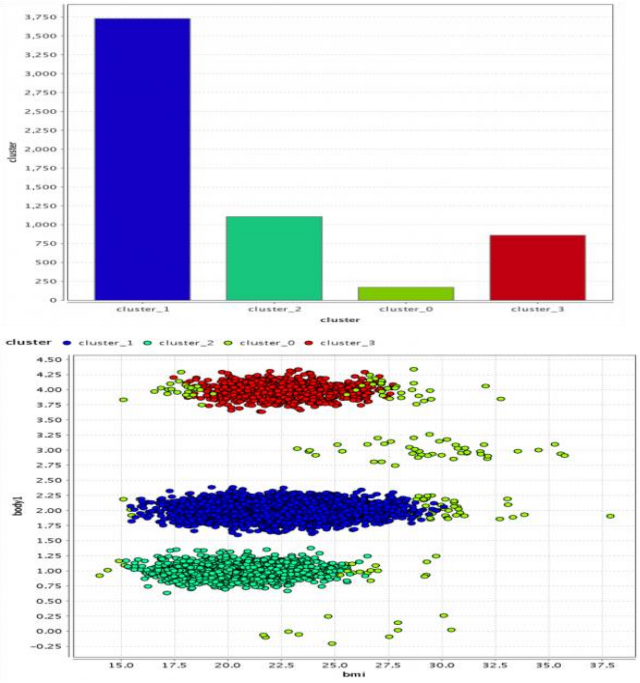


Figure 5- Visualizing the clusters shows us that there are not enough "heavy" people to form a cluster (for body type, Doesn't Matter:0, Slim: 1, Average: 2, Heavy: 3, Athletic: 4). For these plots some scatter has been introduced to make them easier to read, so for example not all the slim profiles are placed horizontally exactly along the 1.0 marker but are dispersed with 1.0 as the center.

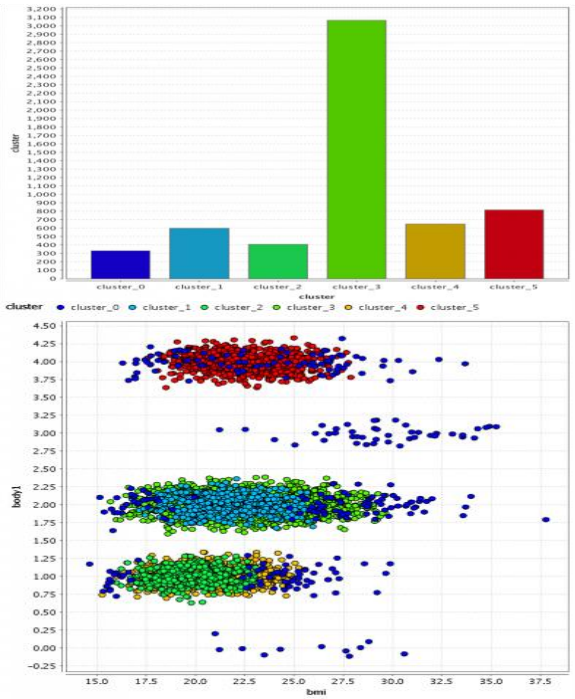


Figure 6 - Visualizing them shows that there are three male clusters and two female clusters. There is a male and female slim cluster (body type = 1.0), a male and female average cluster (body type = 2.0) and a male athletic cluster (body type = 4). There is no female athletic cluster and there are no heavy clusters (body type = 3.0) whatsoever.

4.2 Visualization of dowry cases registered in India using data visualization technique:-

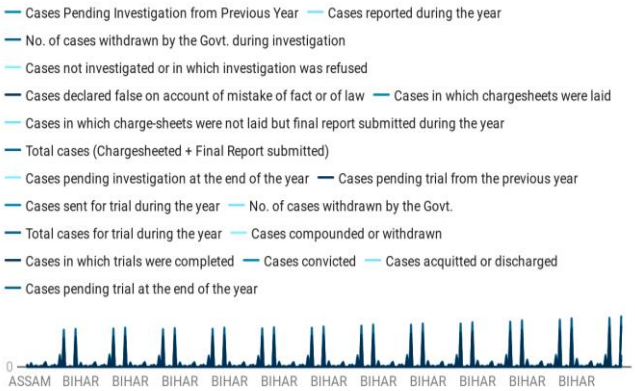


FIGURE 7: CASES REGISTERED

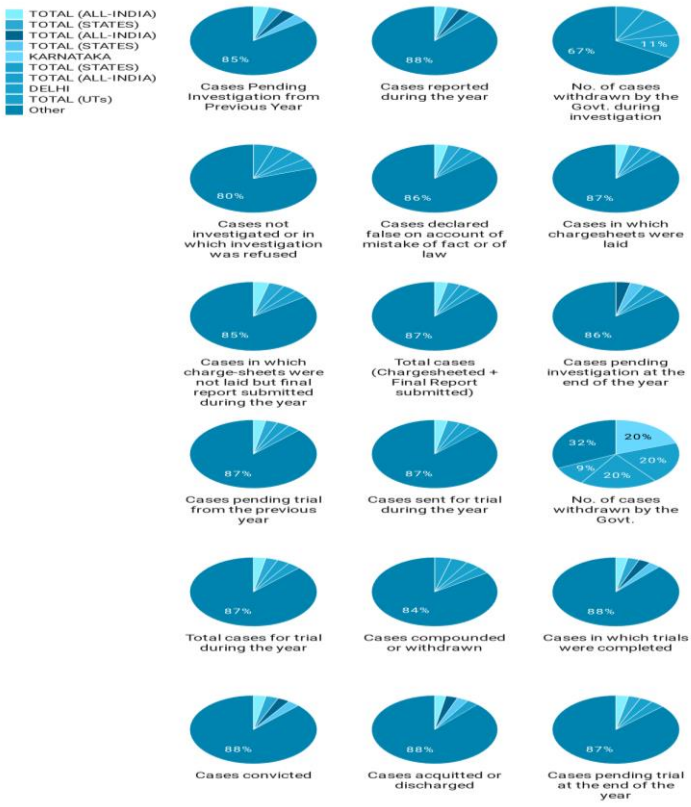


Figure 8: MULTIPLE PIES

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.722-728, October 2020

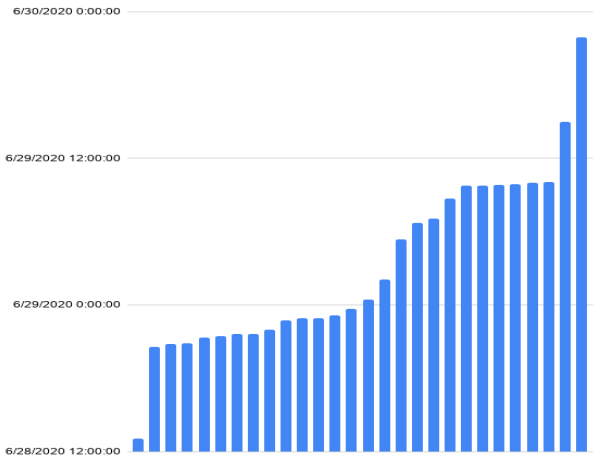
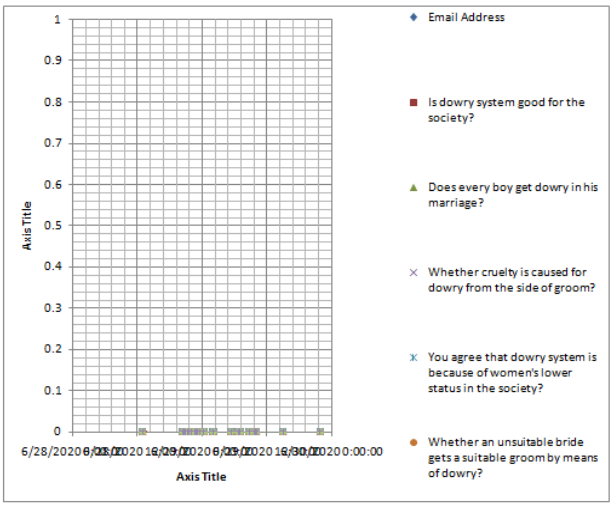
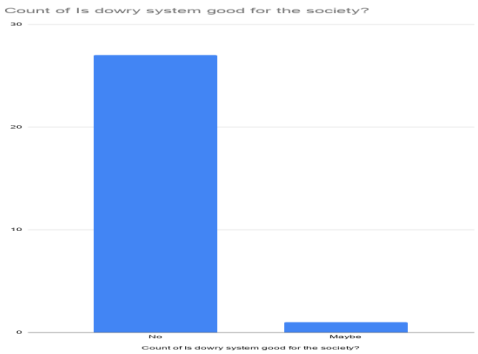


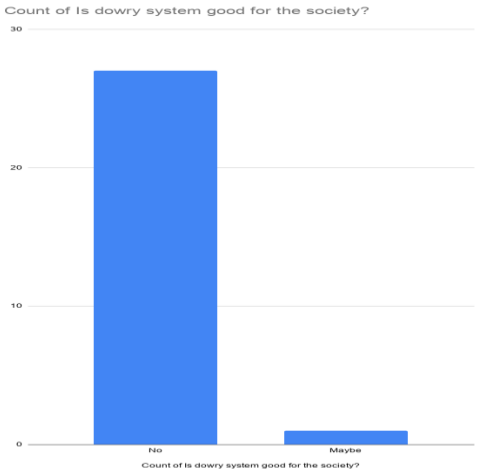
Figure 9: SCATTERED POINT GRAPH

V. SURVEY ANALYSIS

For this survey analysis we have created a Google form and recorded the responses of too many individuals. We have created a questionnaire compressing the too many questions that related to the peoples view overcast system analysis and dowry.







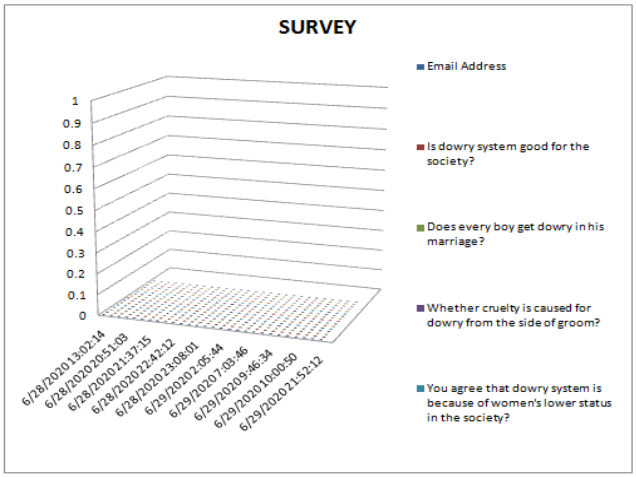


Figure 10 – DEMENSION OF RESPONSE

VI.CONCLUSION

In this paper we have used too many data sets on the Matrimonial website. We have used several clustering in machine learning algorithms to predict the horoscope preferences through the data sets and collected by opinion mining and sentiment analysis of the data on the Matrimonial website. ImplementedK - mean clustering for visualizing some cluster patterns out of the data. We have seen the total number of Dowry cases in India and their visualization. We have undergone a service data collection in research through which we have expected some out point of views of the People's over the Dowry and caste system analysis.

REFERENCES

[1] Bologa, A. R., Bologa, R., & Florea, A. (2013). Big data and specific analysis methods for insurance fraud detection. *Database Systems Journal*, *4*(4), 30-39.
[2] Bajpai, A., & Dayanand, A. A. (2018). Big Data Analytics in Cyber Security.
[3] Sudha, C., & Akila, D. (2019). Detection OFAES Algorithm for Data Security on Credit Card Transaction.
[4] Artís, M., Ayuso, M., & Guillen, M. (1999). Modelling different types of automobile insurance fraud behaviour in the Spanish market. *Insurance: Mathematics and Economics*, *24*(1-2), 67-81.
[5] Abbott, D. W., Matkovsky, I. P., & Elder, J. F. (1998, October). An evaluation of high-end data mining tools for fraud detection. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)* (Vol. 3, pp. 2836-2841). IEEE.
[6] Barse, E. L., Kvarnstrom, H., & Jonsson, E. (2003, December). Synthesizing test data for fraud detection systems. In *19th Annual Computer Security Applications Conference, 2003. Proceedings.* (pp. 384-394). IEEE.
[7] Bell, T. B., & Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory*, *19*(1), 169-184.
[8] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, 235-255.
[9] Rutrell, Y. (2012). Analytics platform helps agencies fight cyber crime, government computer news.
[10] Lakshmi, T. M., Martin, A., Begum, R. M., & Venkatesan, V. P. (2013). An analysis on performance of decision tree algorithms using student's qualitative data. *International journal of modern education and computer science*, *5*(5), 18.
[11] Lavanya, D., & Rani, K. U. (2011). Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications*, *26*(4), 1-4.
[12] Bhukya, D. P., & Ramachandram, S. (2010). Decision tree induction: an approach for data classification using AVL-tree. *International Journal of Computer and Electrical Engineering*, *2*(4), 660.
[13] Aher, S. B., & Lobo, L. M. R. J. (2011, March). Data mining in educational system using weka. In *International Conference on Emerging Technology Trends (ICETT)* (Vol. 3, pp. 20-25).
[15] Kumar, S. A. (2011). Efficiency of decision trees in predicting student's academic performance.
[16] http://iacs-courses.seas.harvard.edu/ courses/iacs_projects /matrimony_data_exploration/meta-analysis.html
[17]https://data.world/datasets/marriage