Prediction Analysis of the Primary Health Conditions of the lower Strata Community using Machine Learning

Pooja Singh, Seema Shokeen, Kriti, Kajol Gupta

Abstract Good health acts as a catalyst in realizing the individual's capabilities and thus contributes to the well- being of the society. It has been observed as a psychology that healthy people are proved to be more productive and contributes more towards the economic development for any nation. On the contrary the ill-health people may not stand up to the mark of the complete realization of their psychological, social and economic capabilities, and hence have to bear the financial implications in terms of loss of income. Productivity and attainment of a good medical care. There are many factors that provide these poor health conditions. This paper consist of all these primary factors and total evaluation of their health by which we can predict their overall health. As the health is directly proportional to the productivity. So as, people can also concentrate on their health. The main aim of this paper is to provide a analysis on the research field of healthcare data. In this paper, we focus on the database which was collected in form of questionnaire. These data are then analysed using a machine learning classifier with and without feature subset selection(monthly expenditure, family income, food consumed, Age of the person, source of drinking water, frequently faced illnesses, condition of living, etc) separately for prediction of a person's overall health.

Keywords

Data mining, Healthcare, Prediction analysis, Features selection, Classification.

I. INTRODUCTION

Delhi being the capital city of India is very prestigious for the country in many aspects. It is the place where the all major the decisions regarding government policies and schemes takes place which is to be made applicable to the whole nation. But still there are so many sectors which are still lagging behind which requires a necessary look about. Health sector of the state is one among those affected domains of the country. There has been a major shortfall and a big disparity seen in the rural and the urban health of the state, where the urban health of the people is still better as compared to the rural population.

To bridge this gap several remarkable steps have been taken by the government in this direction.

The major step taken is the launch of National Rural Health Mission (NRHM) in 2005, which has the prime objective to significantly improve the rural health care delivery by using the techniques of planning and decision-making and several initiatives taken for financial risk protection.

It also initiated Rashtriya Swasthya Bima Yojana, the milestone of Central Government, which provides cashless facility for hospitalization to the poor

people population. Approximately 35 million families have been benefited by this scheme nationwide till date.

A health related survey was conducted in Delhi for 431 families of urban poor in form of questionnaire, The questionnaire of the study comprises of questions that are arranged in various categories like Personal details, Educational Qualification, Personal Health Condition, Hospitals and Clinics Facilities and the household Condition of the respondents, to obtain various significant information to check their overall health conditions and the gaps in the various domains which obstructs the good health care facilities to be delivered to the respondents. The data collected for all the above mentioned parameters have been analysed to observe several kinds of results that are demonstrated. The prime motive to conduct the survey is to identify certain steps which may be taken in the positive direction to improve the current health situations of the people living in these areas.

Data acquired in this manner, can be used for predictive analysis process.Data mining is the process of analysing data from various perspectives and summarizing it into information. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified.Data mining typically involves developing rules for classifications based on features in the training set, which is used to classify future data and develop a better understanding of each class (which have been classified) in the database.

II. STATISTICAL ANALYSIS OF OVERALL HEALTH OF THE RESPONDENTS

The health surveys are conducted in form a structured questionnaire, which is having personal as well as health related questions, from the lower communities living in JJ colonies of Dwarka, New Delhi. Then these data collected from the survey was digitized and statically analysed using SPSS tools.

Revised Manuscript Received on 30 May 2019.

***** Correspondence Author

Pooja Singh*, Department of Computer Science, Maharaja Surajmal Institute, Affiliated to GGSIPU, New Delhi, India

Seema Shokeen, Department of Business Administration, Maharaja Surajmal Institute, Affiliated to GGSIPU, New Delhi, India

Kriti, Department of Computer Science, Maharaja Surajmal Institute, Affiliated to GGSIPU, New Delhi, India

Kajol Gupta, Department of Computer Science, Maharaja Surajmal Institute, Affiliated to GGSIPU, New Delhi, India

III. FACTORS AFFECTING THE OVERALL HEALTH OF AN INDIVIDUAL

From the analysis we get some factors which are affecting the overall health of an individual. These factors are listed below.

A. Frequently Faced Problems

Table I: Frequently Faced Problems

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	NA	43	10.0	10.0	10.0
	Fever	136	31.6	31.6	41.5
	Flu	57	13.2	13.2	54.8
	Cold and Cough	91	21.1	21.1	75.9
	Stomach Ache	30	7.0	7.0	82.8
	Muscular Pain	27	6.3	6.3	89.1
	Headache	30	7.0	7.0	96.1
	Skin Problem	10	2.3	2.3	98.4
	Breathing Problem	7	1.6	1.6	100.0
	Total	**431**	**100.0**	**100.0**	

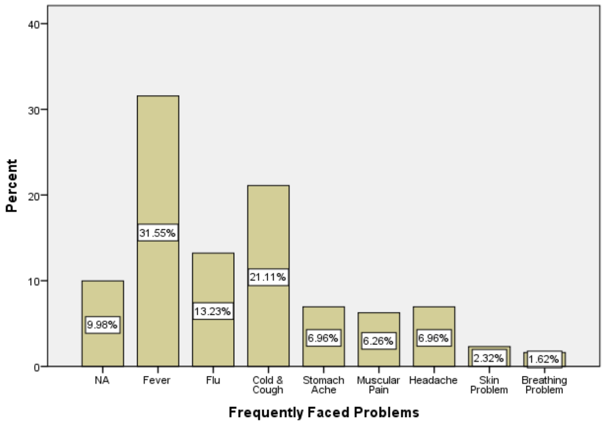

Fig. I: Frequency Distribution plot of the Frequently Faced Problems of the survey

This table depicts the frequency of the most frequently faced problems of the people in the survey. The data indicated very clearly that around one third of the population (31.55%) suffered from a very common problem of fever. Approximately one fifth (21.11%) of the population under survey suffered from common cold and cough. 13.23% of the population suffered from the problem of flu. Around 6% - 7% of the total surveyed population suffered from various kinds of pains such as stomach ache, headache and muscular pain. And roughly a total of 4% (2.32% + 1.62%) suffered from skin problem and breathing problem. At the end there were 9.98% of the people who were not applicable for the same. The above data clearly depicts that the people suffered from very common primary ailments and not from any big or major health issues, which shows a need of much more and better primary health care facilities.

B. Family Monthly Income

Table II: Family monthly income

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	below 5K	30	7.0	12.1	12.1
	5K-10k	85	19.7	34.3	46.4
	10k-15k	72	16.7	29.0	75.4
	Above 15k	61	14.2	24.6	100.0
	Total	**248**	**57.5**	**100.0**	
Missing	System	183	42.5		
Total		**431**	**100.0**		

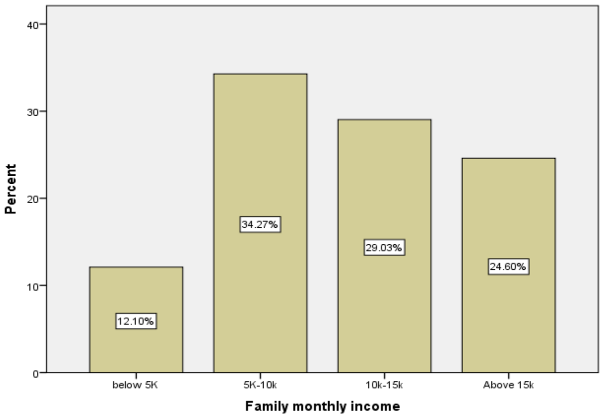

Fig. II: Frequency Distribution plot of the survey monthly income

The survey conducted indicated that approximately one-fourth of the total population surveyed i.e. 24.60% is having a monthly family income more than Rs.15, 000. 29.03% of the population is earning in the range of Rs.10, 000 – Rs. 15,000 monthly, 34.27% of the population is earning somewhere between Rs. 5, 000 – Rs. 10, 000 and 12.10% of the total surveyed population is still earning below Rs.5, 000 monthly which makes it very difficult for a family to maintain a healthy standard of living.

C. Monthly Expenditure on Health

Table III: Monthly Health Expenditure

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	NA	17	3.9	3.9	3.9
	Below 1K	183	42.5	42.5	46.4
	1K-2K	151	35.0	35.0	81.4
	2k-3k	49	11.4	11.4	92.8
	3K or above	31	7.2	7.2	100.0
	Total	**431**	**100.0**	**100.0**	

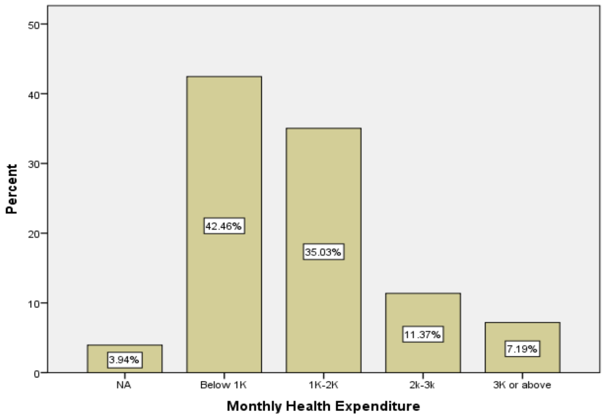

Fig. III: Frequency Distribution plot of the Monthly Health Expenditure of the survey

The survey showed that roughly half of the population under the survey i.e. 42.46% spends less than Rs.1000 monthly on their medical expenses. 35.03% of the people spent between Rs.1000 to Rs.2000 monthly as their medical expenditure. 11.37% people spent Rs.2000 to Rs.3000 monthly on their health expenses. A very small percentage of the population under survey, 7.19% spent more than Rs.3000 monthly on their health. The survey also depicted that 3.94% of the people did not have any fixed budget for their health expenditure and that is also the reason that they suffer a bad primary health.

D. Hospital Visit

Table IV: Hospital Visits of Survey Group

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	NA	25	5.8	5.8	5.8
	Once a Week	74	17.2	17.2	23.0
	Twice a Week	82	19.0	19.0	42.0
	Monthly	139	32.3	32.3	74.2
	None	111	25.8	25.8	100.0
	Total	**431**	**100.0**	**100.0**	

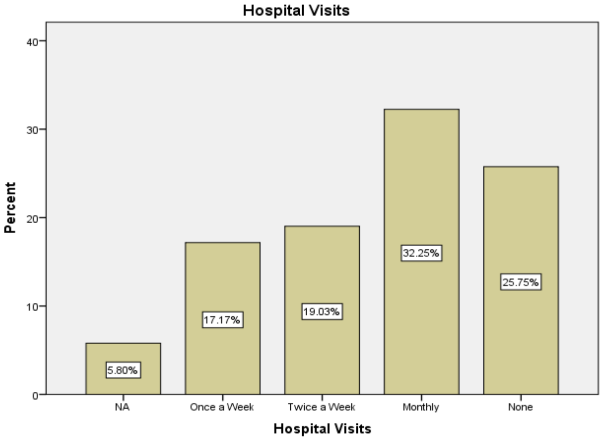

Fig. IV: Frequency Distribution plot of the survey data showing hospital visits

Above table depicts the frequency of the visits to the hospitals of the people in the survey, survey indicated that almost one third (32.25%) of population under survey visited hospital monthly and about one fifth (19.03%) visited the hospital twice a week and 17.17% utilized the hospital facility only once a week. With quarter of population never visiting the hospital indicating good health, around 5.8% did not answer this question. Also this is to be noted that around 68.25% (17.17% + 19.03% + 32.25% = 68.25%) usually visits the hospital which indicates poor health in survey data.

E. Overall health

Table V: Overall Health

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	NA	25	5.8	5.8	5.8
	Good	118	27.4	27.4	33.2
	Mild	125	29.0	29.0	62.2
	Moderate	138	32.0	32.0	94.2
	Severely ill	19	4.4	4.4	98.6
	Chronic	6	1.4	1.4	100.0
	Total	**431**	**100.0**	**100.0**	

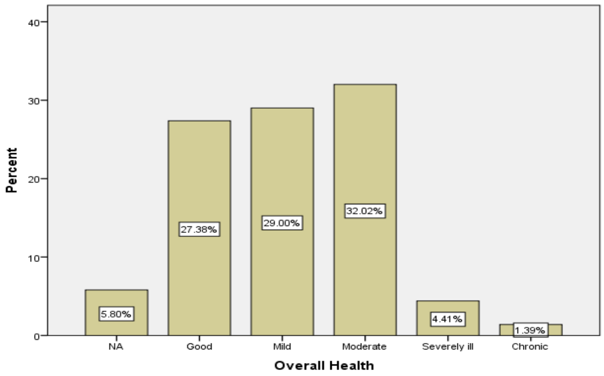

Fig. V: Frequency Distribution plot of the overall health of the survey conducted

Above table depicts the overall health of the respondents of the survey. Around one third of the population (32.02%) is just having moderate health, 29% of the population is having a very mild health conditions and only 27.38% of the population claimed a good. 4.41% of people were severely ill and only 1.39% people suffered chronic conditions. This shows that the majority of the population is having a moderate health which is a sign of deteriorated primary health facilities.

IV. PREDICTIVE ANALYSIS OF THE HEALTHCARE DATA

We are living in a world which is completely surrounded by a very large amount of data that is collected either on the daily basis, monthly basis or annual basis. This data can become a very useful tool if utilized appropriately for making several decisions [26]. Here comes the need of DataMining which helps in bringing this collected data into a useful information**.** According toKamber, data mining [9] is defined as the process of knowledge discovery from data.

Predictive analysis is the technique of extracting information, trends and behaviors from the existing historical data sets in order to determine patterns and predict the outcomes and trends, as discussed by Bellazi[4]. It provides the suggestion for the expected future outcomes that helps in doing the decision making for several kinds of trends in the existing dataset. The predictive data analytics utilizes a huge variety of statistical modelling, data mining, and machine learning techniques to study a large amount of current and historic data, which further allows the analysts to make several useful and exciting predictions about the future [6] . This future prediction becomes a very useful tool for taking several required decisions.

V. TOOL INVOLVED IN PREDICTIVE ANALYSIS OF HEALTHCARE DATA

SVM, Support Vector Machine defines decision boundaries and is are based on the concept of decision planes. A decision plane is one that distinguish a set of objects belonging from different classes. It is a classifying methodology that performs classification tasks by creating hyperplanes in a multidimensional space that separates cases of various class labels. SVM supports regression as well as classification tasks and can handle multiple categorical variables [24].

VI. STEPS INVOLVED IN PREDICTIVE ANALYSIS OF HEALTHCARE DATA

The health care data set gathered from the survey conducted; contain too many features. The process ofFeature Extraction is defined as identification and elimination of irrelevant, weakly relevant or may be completely redundant attributes or dimensions that exists in the data set provided for performing the analysis. This technique will help in the identification of the minimal subset of attributes which are appropriate enough in resulting probability distribution of data classes that is nearly similar to the original distribution of the attributes obtained using all the attributes.

From these identified features we needed to find out the most important features or attributes to identify the choice of preferred hospital by respondents. To Check the performance of Classification vs. the Feature Selection (FS) based classification we first ran Classification, using support vector classifier and found its accuracy which was around 52.90% however when the Feature Subset Selection (FSS) [2] was applied the classification accuracy rose up to 80.00% and the main features or identifiers of selectors were found to be Monthly Health Expenditure, Chronic Condition, Family income, food consumed, Age, Source of drinking water, Frequently faced problems, Hospital visit, Living condition.

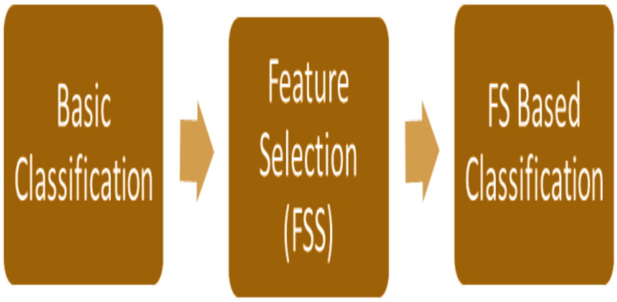

Fig. VI: Steps of Feature Selection Based Classification

Then data cleaning was performed which involves filling missing values, removing rows, reducing data size to create a error free dataset. So, after cleaning of data set we have 423 sets of data. Then we will split the dataset into two, 80% of which we will use to train our model i.e. 338 datasets and 20% of that we will hold back as for validating i.e. 85 datasets. The results of the prediction analysis is shown in form of Confusion Matrix.

A **confusion matrix** is a technique for summarizing the performance of a classification algorithm. Composing a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making .It is a very useful tool for analysing how well your classifier can recognize tuples of different classes (TP, TN, FP, FN) where TP and TN tells us when the classifier is getting the things right while FP and FN tells us when the classifier is getting the things wrong [9].

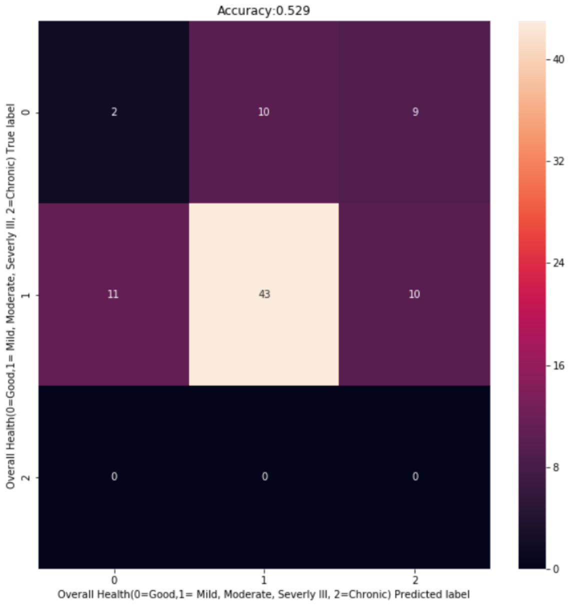

Fig. VII: Confusion Matrix 1 before applying Feature Subset Selection

A. Interpretation of Classification Result before feature subset selection

The True value of overall health of 24.70% ((2+10+9)*100/85) of validating dataset is good and 9.52% (2*100/(2+10+9)) of it is also predicted correctly. Whereas, 90.47% ((10+9)*100/(2+10+9)) of it is predicted incorrect. Then the true value of overall health of 75.29% ((11+43+10)*100/85) of total validating data set is Mild and 67.18% (43*100/(11+43+10)) of it is also predicted correctly. Whereas, 32.81% ((11+10)*100/(11+43+10) of it is predicted incorrectly by the classifier.

B. Accuracy before applying FSS

(2+43+0)/(2+10+9+11+43+10) = 45/85 = 52.94%

C. Error rate before applying the classifier

100 – 52.94 = 47.05%

(10+9+11+10)/(2+10+9+11+43+10) = 40/85 = 47.05%

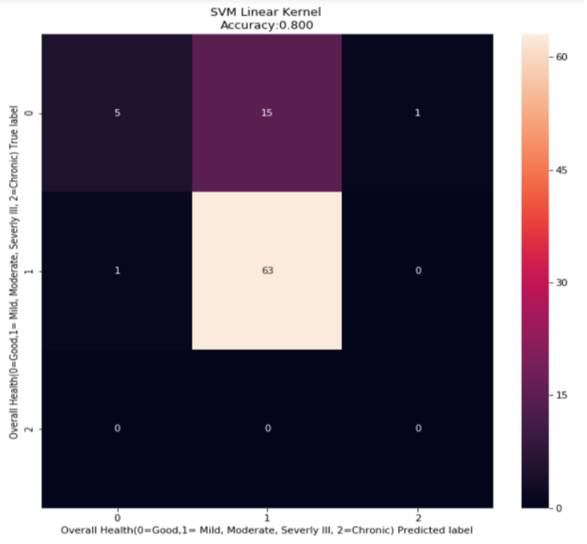

Fig. VIII: Confusion Matrix 2 after applying Feature Subset Selection

D. Interpretation of Classification Result after feature subset selection

The True value of overall health of 24.70% (((5+15+1)*100/85) of validating dataset is good and 23.80% (5*100/(5+15+1)) of it is also predicted correctly. Whereas, 76.19% ((15+1)*100/(5+15+1)) of it is predicted incorrect. Then the true value of overall health of 75.29% ((1+63)*100/85) of total validating data set is Mild and 98.43% (63*100/(63+1)) of it is also predicted correctly. Whereas, 1.56% (1*100/(63+1)) of it is predicted incorrectly by the classifier.

E. Accuracy after applying FSS

(5+63+0)/(5+15+1+1+63) = 68/85 = 80.00%

F. Error rate before applying the classifier

100 – 80.00 = 20.00%

(15+1+1)/(5+15+1+1+63) = 17/85 = 20.00%

Table VI: accuracy after and before applying FSS

	Before FSS	**After FSS**
Correct prediction	45	68
% of correct prediction	52.94%	80.00%
Incorrect prediction	40	17
% of incorrect prediction	47.05%	20.00%

With reference from Table 7, we can see very clearly that after applying the feature selection method there has been a significant increase in the accuracy of the model. The accuracy increased by approximately 27% (80.00-52.94) after applying FSS in the model.

We can also see that there has been a significant decrease in the error of the model. The error decreased by approximately 27% (47.05-20.00) after applying FSS in the model.

VII. CONCLUSION

The accuracy before applying FSS was 52.94% and the error rate was 47.05% whereas after applying FSS in classifier accuracy increased by approx 27% (80.00-52.94) and error rate is decreased by approx 27% (47.05-20.00). From the above observation, we concluded that the prediction of the overall health of an individual with the help of machine learning classifier with feature subset selection is more efficient than without Feature subset classifier which means after applying FSS the overall health predicted by the classifier will be more accurate and error rate will be low as compared to without applying FSS in the classifier.

Using the results which we have obtained from the prediction analysis we can also improve their overall health conditions by considering only the major factors which are affecting the health condition of an individual. The overall health of people are interdependent on each other. This study is used to find major aspects in the health of people of a particular area, how their livelihood affects their health and vice-versa. So, from this study, we can create awareness among individuals who are not financially strong and it can also improvise their health.

REFERENCES

1. Acharya, A., and Ranson, M.K., Health Care Financing for the Poor: Community-Based Health Insurance Schemes in Gujarat. Economic and Political Weekly, 2005, 40(38), pp. 4141-4150.
2. Anbarasi, M., Anupriya, E. and Iyengar, N. C. S., Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology, 2010, 2(10), pp. 5370–5376.
3. Anderson Graduate School of Management. (2012). Data Mining: What is Data Mining? Retrieved from UCLA: Anderson Graduate School of Management: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm
4. Bellazzi, R, and Zupan, B., Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics, 2008, 77(2), pp. 81–97.
5. Delhi Human Development Report, Health and Health care, 2013.
6. Ganjir, V., Sarkar, B. K., and Kumar, R., Big data analytics for healthcare. International Journal of Research in Engineering, Technology and Science, 2016, pp. 1-6
7. Golechha, M., Healthcare agenda for the Indian government. Indian Journal of Medical Research, 2015, 141(2), pp. 151-153.
8. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.
9. Han, J., Kamber, M. and Pie, J., Data Mining: Concepts and Techniques (Third Edition), Elsevier Publishers, 2007.
10. Indian Public Health Standards (IPHS) Guidelines for Primary Health Centres Revised, Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India, 2012.
11. K, S., Rani, K. B., & A, G. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 251-252.
12. Lahariya, C., Mohalla Clinics of Delhi, India: Could these become platform to strengthen primary healthcare? Journey of Family Medicine and Primary Care, 2017, 6(1) pp. 1-10.
13. Ministry of Health and Family Welfare, Framework for Implementation- National Health Mission (2012-17). http://pib.nic.in/newsite/PrintRelease.aspx?relid=159404
14. Ministry of Health and Family Welfare, Rural Health Statistics 2014-15. http://wcd.nic.in/sites/default/files/RHS_1.pdf
15. Ministry of Health and Family Welfare, Annual report 2015-16, Information, Education and Communication. https://mohfw.gov.in/sites/default/files/17563256478856633221.pdf
16. Muni Kumar, N., and Manjula, R., Role of Big Data Analytics in Rural Health Care - A Step towards SvasthBharath. International Journal of Computer Science and Information Technologies (IJCSIT), 2014, 5(6), pp. 7172–7178.
17. National Family Health Survey – 4, India fact sheet. Ministry of Health and Family Welfare, Government of India. International Institute for Population Sciences, 2015-16.
18. Peters, D. H., Garg, A., Bloom, G., Walker, D. G., Brieger, W. R., and Rahman, M. H., Poverty and Access to Health Care in Developing Countries. New York Academy of Sciences. 2008, 1136, pp. 161–171.
19. Raghupathi, W., and Raghupathi, V., Big data analytics in healthcare: promise and potential. Health information science and systems, 2014, 2(1).
20. Rasheed, N., Arya, S., and Acharya, A., Client satisfaction and perceptions about quality of health care at a primary health center of Delhi, India. Indian Journal of Community Health, 2012, 24(3), pp. 237–242.
21. Rural Health Statistics Report. Ministry of Health and Family Welfare, Statistics Division, Government of India, 2014-15.
22. Srivastava, S., Pant, M., Abraham, A., and Agrawal, N., The Technological Growth in e-Health Services.Computational and mathematical methods in medicine, 2015.
23. State of Urban Health in India, Ministry of Health and Family Welfare. http://www.pbnrhm.org/docs/nuhm_framework_implementation.pdf
24. Support Vector Machines (SVM). http://www.statsoft.com/textbook/support-vector-machines
25. The Clinical Establishments (Registration and Regulation) Act, LokSabha Bill, 2010.
26. Tomar, D. and Agarwal, S., A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 2013, 5(5), pp. 241–266.
27. World Health Organization, Health and Development, Poverty and Health. http://www.who.int/hdp/poverty/en/
28. Your First Machine Learning Project in Python Step-by-Step. https://machinelearningmastery.com/machine-learning-in-python-step-by-step/