INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT

IJRCM

A Monthly Double-Blind Peer Reviewed (Refereed/Juried) Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Cabell's Directories of Publishing Opportunities, U.S.A., Google Scholar, Indian Citation Index (ICI), J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)], Index Copernicus Publishers Panel, Poland with IC Value of 5.09 (2012) & number of libraries all around the world.

Circulated all over the world & Google has verified that scholars of more than **6185 Cities** in **195 countries/territories** are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

http://ijrcm.org.in/

CONTENTS

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to the recent developments & practices in the areas of Computer Science & Applications; Commerce; Business; Finance; Marketing; Human Resource Management; General Management; Banking; Economics; Tourism Administration & Management; Education; Law; Library & Information Science; Defence & Strategic Studies; Electronic Science; Corporate Governance; Industrial Relations; and emerging paradigms in allied subjects like Accounting; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Rural Economics; Co-operation; Demography: Development Planning; Development Studies; Applied Economics; Development Economics; Business Economics; Monetary Policy; Public Policy Economics; Real Estate; Regional Economics; Political Science; Continuing Education; Labour Welfare; Philosophy; Psychology; Sociology; Tax Accounting; Advertising & Promotion Management; Management Information Systems (MIS); Business Law; Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labour Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; International Relations; Human Rights & Duties; Public Administration; Population Studies; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism & Hospitality; Transportation Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic; Web Design and emerging paradigms in allied subjects.

Anybody can submit the **soft copy** of unpublished novel; original; empirical and high quality **research work/manuscript** ***anytime*** in ***M.S. Word format*** after preparing the same as per our **GUIDELINES FOR SUBMISSION**; at our email address i.e. infoijrcm@gmail.com or online by clicking the link **online submission** as given on our website (***FOR ONLINE SUBMISSION, CLICK HERE***).

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION**:

DATED: _____________

THE EDITOR

IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF ____.**

(e.g. Finance/Mkt./HRM/General Mgt./Engineering/Economics/Computer/IT/ Education/Psychology/Law/Math/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript titled '____' for likely publication in one of your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published anywhere in any language fully or partly, nor it is under review for publication elsewhere.

I affirm that all the co-authors of this manuscript have seen the submitted version of the manuscript and have agreed to inclusion of their names as co-authors.

Also, if my/our manuscript is accepted, I agree to comply with the formalities as given on the website of the journal. The Journal has discretion to publish our contribution in any of its journals.

NAME OF CORRESPONDING AUTHOR	:
Designation/Post*	:
Institution/College/University with full address & Pin Code	:
Residential address with Pin Code	:
Mobile Number (s) with country ISD code	:
Is WhatsApp or Viber active on your above noted Mobile Number (Yes/No)	:
Landline Number (s) with country ISD code	:
E-mail Address	:
Alternate E-mail Address	:
Nationality	:

* i.e. Alumnus (Male Alumni), Alumna (Female Alumni), Student, Research Scholar (M. Phil), Research Scholar (Ph. D.), JRF, Research Assistant, Assistant Lecturer, Lecturer, Senior Lecturer, Junior Assistant Professor, Assistant Professor, Senior Assistant Professor, Co-ordinator, Reader, Associate Professor, Professor, Head, Vice-Principal, Dy. Director, Principal, Director, Dean, President, Vice Chancellor, Industry Designation **etc.** ***The qualification of author is not acceptable for the purpose***.

NOTES:

a) The whole manuscript has to be in ***ONE MS WORD FILE*** only, which will start from the covering letter, inside the manuscript. ***pdf. version* is liable to be rejected without any consideration**.

b) The sender is required to mention the following in the **SUBJECT COLUMN of the mail**:

New Manuscript for Review in the area of (e.g. Finance/Marketing/HRM/General Mgt./Engineering/Economics/Computer/IT/Education/Psychology/Law/Math/other, please specify)

c) There is no need to give any text in the body of the mail, except the cases where the author wishes to give any **specific message** w.r.t. to the manuscript.

d) The total size of the file containing the manuscript is expected to be below **1000 KB**.

e) Only the **Abstract will not be considered for review** and the author is required to submit the **complete manuscript** in the first instance.

f) ***The journal gives acknowledgement w.r.t. the receipt of every email within twenty-four hours*** and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of the manuscript, within two days of its submission, the corresponding author is required to demand for the same by sending a separate mail to the journal.

g) The author (s) name or details should not appear anywhere on the body of the manuscript, except on the covering letter and the cover page of the manuscript, in the manner as mentioned in the guidelines.

2. **MANUSCRIPT TITLE:** The title of the paper should be typed in **bold letters**, **centered** and **fully capitalised**.

3. **AUTHOR NAME (S) & AFFILIATIONS:** Author (s) **name**, **designation**, **affiliation** (s), **address**, **mobile/landline number** (s), and **email/alternate email address** should be given underneath the title.

4. **ACKNOWLEDGMENTS:** Acknowledgements can be given to reviewers, guides, funding institutions, etc., if any.

5. **ABSTRACT:** Abstract should be in **fully Italic printing**, ranging between **150** to **300 words**. The abstract must be informative and elucidating the background, aims, methods, results & conclusion in a **SINGLE PARA**. ***Abbreviations must be mentioned in full***.

6. **KEYWORDS:** Abstract must be followed by a list of keywords, subject to the maximum of **five**. These should be arranged in alphabetic order separated by commas and full stop at the end. All words of the keywords, including the first one should be in small letters, except special words e.g. name of the Countries, abbreviations etc.

7. **JEL CODE:** Provide the appropriate Journal of Economic Literature Classification System code (s). JEL codes are available at www.aeaweb.org/econlit/jelCodes.php. However, mentioning of JEL Code is not mandatory.

8. **MANUSCRIPT:** Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. *It should be free from any errors i.e.* ***grammatical, spelling*** *or* ***punctuation.*** *It must be thoroughly edited at your end*.

9. **HEADINGS:** All the headings must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

10. **SUB-HEADINGS:** All the sub-headings must be bold-faced, aligned left and fully capitalised.

11. **MAIN TEXT:**

THE MAIN TEXT SHOULD FOLLOW THE FOLLOWING SEQUENCE:

INTRODUCTION
REVIEW OF LITERATURE
NEED/IMPORTANCE OF THE STUDY
STATEMENT OF THE PROBLEM
OBJECTIVES
HYPOTHESIS (ES)
RESEARCH METHODOLOGY
RESULTS & DISCUSSION
FINDINGS
RECOMMENDATIONS/SUGGESTIONS
CONCLUSIONS
LIMITATIONS
SCOPE FOR FURTHER RESEARCH
REFERENCES
APPENDIX/ANNEXURE

The manuscript should preferably be in ***2000 to 5000 WORDS***. But the limits can vary depending on the nature of the manuscript.

12. **FIGURES & TABLES:** These should be simple, crystal **CLEAR**, **centered**, **separately numbered** & self-explained, and the **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. *It should be ensured that the tables/figures are referred to from the main text*.

13. **EQUATIONS/FORMULAE:** These should be consecutively numbered in parenthesis, left aligned with equation/formulae number placed at the right. The equation editor provided with standard versions of Microsoft Word may be utilised. If any other equation editor is utilised, author must confirm that these equations may be viewed and edited in versions of Microsoft Office that does not have the editor.

14. **ACRONYMS:** These should not be used in the abstract. The use of acronyms is elsewhere is acceptable. Acronyms should be defined on its first use in each section e.g. Reserve Bank of India (RBI). Acronyms should be redefined on first use in subsequent sections.

15. **REFERENCES:** The list of all references should be alphabetically arranged. ***The author (s) should mention only the actually utilised references in the preparation of manuscript*** and they may follow Harvard Style of Referencing. **Also check to ensure that everything that you are including in the reference section is duly cited in the paper**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc., in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italic printing. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parenthesis.
- ***Headers, footers, endnotes** and **footnotes** should **not be used** in the document.* However, **you can mention short notes to elucidate some specific point**, which may be placed in number orders before the references.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–23

UNPUBLISHED DISSERTATIONS

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, Kurukshetra University, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITES

- Garg, Bhavet (2011): Towards a New Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

AN APPROACH TOWARDS A COMPREHENSIVE BIG DATA ALGORITHM IN DEDUCING AND RESOLVING HEALTHCARE PATTERNS IN INDIA'S RURAL-POOR

MEGHA PANJWANI
STUDENT
BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI

PRATEEK BAJAJ
STUDENT
BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI

ABSTRACT

With India becoming one of the world's fastest growing economies, there is still an extremely large part of the country lying under the poverty index, with more than 250 million citizens still marked as poor. Adding to that, statistics from the World Bank Organization suggest that more than 60 million children in the country are malnourished; with about 45% stunted (too short for their age), 20% wasted (too thin for their height, indicating acute malnutrition), 75% anemic, and 57% Vitamin A deficient. These statistics lead to an incriminating look into how the country's rural poor are reeling with health-care and nutrition needs. With technology advancing to greater heights in the field of data analytics, this paper aims at using big data analytics concepts such as data munging, visualization, clustering, and the likes into developing a single algorithm that would establish correlation between a person's nutrition and day-to-day activities with their health care patterns, to provide better healthcare predictions.

KEYWORDS

big data, health, statistics, analysis, nutrition, patterns, malnourishment, data science, analytics, algorithm.

JEL CODES

I10, C80.

INTRODUCTION

The major aim for the healthcare industry in the world today is to work on value-based business, as compared to the volume-based business that has been present for centuries at a stretch. This requires working on techniques and solutions that add more value to the healthcare practices of today.
The major requirement for such a shift is partly because of the abundance of data available today in the healthcare industry.
Over the past few years, health care data have become much more complex, partly due to the abundance of data available and in part due to the advancement in technology to help collect (or fetch) data from sources that weren't quite feasible a few years back. Due to this, it has been widely believed that healthcare data analytics is an important field of study in order to manage such large amounts of data.
But even with the widespread availability of resources to collect data and the technology to analyze and use it, one of the major hurdles presented today is to the developing countries *(such as India)* that do not have the resources to implement such advanced mechanisms. High costs due to reduced efficiency of scale, lack of proper technology and network availability in rural areas and small clinics *(Bhattarcharjee, et al., 2007).*
In countries such as India, one of the major hurdles being faced today in the healthcare industry is the lack of quality data. If the healthcare industry in the country is able to work on ways to understand healthcare patterns in the major population of India (especially the rural-poor, that constitute an extremely large part of the country), data scientists and analysts along with healthcare researchers and professionals could then use that data to deduce results and solutions to a lot many problems presented today in the country. A few of the major problems in the country today in lieu with the healthcare (directly or indirectly) include *lack of access to proper sanitation in rural areas*—less than 50% rural population with access to improvised sanitation*; lack of total health-care spending*—less than 4.1% of the GDP being spent on healthcare in India; *low number of health-care professionals in the country*—up-to 2 per 1000 persons as compared to up-to 12 per 1000 persons in developed countries such as Australia; *lack of per capita health care spending*—less than 100 USD (Joumard, Kumar, 2015).
This number is even lower in the rural parts and states of the country. States such as Bihar (1 health worker per 1000), Rajasthan (1.4), Uttar Pradesh (1.3), Assam (1.4) show that the plight of number of health workers is also a huge detriment to improving the health conditions of the country. These statistics show a huge gap between developments in the healthcare industry in developing countries such as India and other developed nations.

OBJECTIVES OF THE STUDY

The major objective of the study was to understand the current health state of the rural-poor population specifically in Delhi-NCR region. This particular study further helped to understand the healthy: poor ratio of the surveyed population and give an overview of the health condition based on rating such as 1: Healthy State, 2: Average State, 3: Unhealthy State as an outcome of the classification algorithm applied to the dataset generated after the desired surveyed results were compiled, filtered and grouped as per requirement.
The detailed objectives include:

1) Applying self-predicting algorithm to gather a result set of health condition of rural-poor population classified as either healthy, average or unhealthy.
2) Studying the current health scenarios of the lower poverty line people and using the same data for further predictive and solution-based algorithms.
3) Ultimately, these survey findings can help in policy formation as well as betterment of health conditions of the underprivileged in alignment with other sections of the society.

RESEARCH METHODOLOGY OF THE STUDY

TARGET POPULATION

The target population here are the families belonging to lower income groups. People aged 18+ have been interviewed about the overall conditions of themselves as well as their families. The major areas covered are Subhash Nagar slums in Delhi and J.J. Colony in Dwarka, New Delhi (NCR). Other people that have been interviewed include house maids, security guards, washer-men/washer-women, auto drivers, sweepers, rickshaw pullers, factory workers and other daily wage earners. A target of 700 interviews has been achieved covering the mentioned groups of people residing in the capital city, Delhi.

METHOD OF INTERVIEW

All survey interviews have been carried out in a face-to-face interactive fashion. The questionnaire being in English, all people being interviewed were made to explain each and every question, asked to relate it with their lifestyle and health condition and this way accurate responses from the interviewees were collected.

DATA SOURCE

The data set used for the findings is Primary Data collected through the means of surveying the target population mentioned in the above section. The data is solely original and not taken as a secondary source of data from any references. The data used clearly specifies the current state of people termed as lower strata and also what are their needs that should be fulfilled to provide them a better and healthier lifestyle. The data consists of information stating their primary health issues that need major focus rather than concentrating on providing facilities and equipment(s) for rather bigger and severe health conditions.

TECHNOLOGY USED

The software (s) needed to analyze the various aspects of the survey are as follows:

Microsoft Excel (preferably 2016 version)

MS-Excel is a very powerful tool of the Microsoft office due to its various features. We can organize your numeric or text data in spreadsheets or workbooks and view it in context through excel which will help us make more informed decisions. As one looks at different configurations, Excel learns and recognizes our pattern and auto-completes the remaining data for you. No formulas or macros required. The Tell Me search feature guides you to the feature commands you need to get the results you are looking for. Excel also performs complex analyses for our data. And it summarizes our data with previews of pivot-table options, so that we can compare them and select the one that tells our story best. Excel can recommend the charts and graphs that best illustrate your data patterns. Quickly preview your options and pick those that present your insights most clearly. We can discover and compare different ways to represent our data and our intents visually. When we see the one that shows our data best, apply formatting, spark lines, charts, and tables with a single click. A set of new modern charts and graphs in Excel 2016 help you to present our data in fresh ways. Excel makes it easy to spot trends and patterns in your data by using bars, colors and icons to visually highlight important values. The new one-click forecasting feature in Excel 2016 creates forecasts on your data series with one click to future trends. In this survey analysis, Excel has been used to create the primary data set by entering all recorded responses in form of a spreadsheet which is further used for extracting patterns.

R Studio

RStudio is an open source tool for performing Data Mining operations. In technical terms, RStudio is a cross-platform integrated development environment (IDE) for the R statistical language. The technology to amass data exceeds our abilities to make use of it. Another cool feature which might be of use to some is the manipulate package for dynamically changing plot parameters. People all over the world are turning to R, an open source statistical language, to make sense of data. Inspired by the innovations of R users in science, education, and industry, RStudio develops free and open tools for R and enterprise-ready professional products for teams to scale and share work. RStudio is an integrated development environment (IDE) for the R programming language. Some of its features include: Customizable workbench with all of the tools required to work with R in one place (console, source, plots, workspace, help, history, etc.). Syntax highlighting editor with code completion. Execute code directly from the source editor (line, selection, or file). Full support for authoring Sweave and TeX documents. Runs on all major platforms (Windows, Mac, and Linux) and can also be run as a server, enabling multiple users to access the RStudio IDE using a web browser.

The major benefit of using RStudio in our survey is that it helped in carrying out the Data Mining for the survey responses we collected from various people. The entire process of mining our desired results included creating a database in the Microsoft Excel spreadsheet, importing a "csv" format of this database in the RStudio and then using the various entries of the dataset to plot histograms, bar graphs and pie charts.

Python

Python is a general-purpose language, which means it can be used to build just about anything, which will be made easy with the right tools/libraries. Professionally, Python is great for backend web development, data analysis, artificial intelligence, and scientific computing. Being a very high-level language, Python reads like English, which takes a lot of syntax-learning stress off coding beginners. Python handles a lot of complexity for you, so it is very beginner-friendly in that it allows beginners to focus on learning programming concepts and not have to worry about too much details.

Also, as a dynamically typed language, Python is really flexible. This means there are no hard rules on how to build features, and you'll have more flexibility solving problems using different methods (though the Python philosophy encourages using the obvious way to solve things). Furthermore, Python is also more forgiving of errors, so you'll still be able to compile and run your program until you hit the problematic part.

NEED FOR ANALYSING HEALTHCARE PATTERNS

There have been many surveys conducted across multiple states in the country to analyse health conditions of people living in rural and semi-rural areas. (Singh, Shokeen, et al., 2017).

Statistics suggest that out of hundreds of people surveyed about their health conditions, that a good 35% people consider them to be in the moderate health category, while a good 25% people feel that their health conditions are mild and could get worse at any time.

Out of the people surveyed, a majority of people suffer from common health concerns such as fever, flu, common cold and cough, and the likes; a good 12.3% people face issues related to breathing and skin diseases. The amount of health concerns in people surveyed suggests that one of the major reasons for the prolonged health issues in people is due to the time lapse between their regular health check-ups. About 74% people suggested that they had not visited a doctor in the past 6 months.

There were some other surprising findings too in concern with the health issues of people surveyed. One of the major outcomes was the vast amount of people without the access to filtered water. The surveys conducted suggested that 58.4% of people did not have access to filtered water and had to survive with untreated and unfiltered sources of one of the basic necessities of human life.

One of the major causes of health-care problems also involve the lack of clean surroundings for children. A whopping 20.5% people believed that their surroundings weren't up to the clean hygienic conditions of livability, while up to 61% people rated their surrounding localities as only moderately clean. Cleanliness (or its lack thereof) is one of the major causes of low health standards in the country. A staggering 59.4% of households in rural areas lacked the basic latrine facilities just a few years back. (Swachh Status Report, 2016).

More so, there have been more than 68% of households in 2012 with no arrangements for garbage disposal.

Factors such as lack of sanitation, garbage disposal, regular health check-ups, cleanliness, purity of water, etc. have been known as direct constituents in the deteriorating health of people.

This suggests the answer to the question about the need for a detailed analysis on the healthcare patterns and needs of the people living in the rural and semi-rural parts of the country.

HEALTHCARE SECTOR & BIG DATA ANALYTICS (PRESENT SCENARIO)

Data Mining (in the field of Big Data) is generally described as the process by which data can be gathered, analysed, and stored so that quality information can be fetched out of the same. (Russom, 2011).

Generally, the first step of data processing is gathering and collecting data. This is done through a variety of mediums, including but not limited to surveys (online and/or offline), questionnaires, internet-connected devices, etc.

Once the data has been collected, the next step mostly is the idea on how/where to store the data. *Big Data Storage and Management* (Alkhatib, Khloei, et al., 2015) is considered as the next step for big data analytics. Next, the data is transformed and classified before being ready for proper analysis.

(A) Predictive Analytics in Healthcare Systems

One of the areas where predictive analytics and in turn data is being highly used is the use of healthcare prediction on reducing further risks to patients and thereafter reducing costs. Predictive analytics use medical histories of patients to evaluate and determine further risks. In turn, it also helps in predicting and understanding future medical treatments.

Parkland Hospital in Dallas, Texas has introduced a health-care prediction system that helps scan patients' details and identifies potential risks and outcomes (Jacob, 2012).

(B) Role of Predictive Analytics in Healthcare

The major role of predictive analytics in the healthcare system has been in making sure that specialists are able to achieve results and outcomes, and predict preventive care solutions to patients. This also allows practitioners to work on treatments to avoid further risks to the patients (Conley, et al,. 2008).

PREDICTIVE MODEL (ALGORITHM) FOR HEALTHCARE ANALYTICS

With data from hundreds of residents of localities in and around the National Capital Region of the country, the aim was to generate a model for healthcare analytics converging around the idea of using statistical information such as age, source of drinking water, foods consumed on a daily basis, cooking fuel used, frequency of visits to the doctor, air ventilation and cleanliness in the society, and linking these with the medical conditions that the people have.

The first major task whilst generating a model for Machine Learning is to understand the kind of learning methodology to be followed.

The following explains (in detail) in a step-by-step process the creation of a Machine Learning Methodology (or an algorithm) that can be used in analysing healthcare data and statistics to generate predictive outcomes.

The first step in starting with machine learning on healthcare data is to understand what the data are, whether the data provided to the model are fully functional and can be properly worked on, and making sure that the data are ready to be processed and provided to the Machine Learning Model. In technical terms, this process is known as *Data Preprocessing.*

Data Preprocessing constitutes a variety of steps. The foremost of them all constitutes Data Slicing/Dicing. *Data slicing*, as the name suggests requires breaking down the data into several smaller parts so as to examine it from different viewpoints and understand it better.

For scenarios related to healthcare data, there are fields of information (or labels/columns) which aren't directly relevant for our use case. In such cases, the data scientist's first job is to demarcate the fields most relevant to our analysis.

Once the data slicing is carried out successfully, the next task that needs to be completed falls under *Treating Missing Values.* This involves marking the empty value fields, filling them using one of the many algorithmic techniques available, and then filtering data based on any missing information.

Next comes *Removing Duplicate Values.* This might not seem as an extremely crucial step in solving the problem, but it does weigh in to the end result, as it helps in removing any skewness that might come due to farcical or duplicate values.

One of the most important factors while preprocessing data is making sure that all the values provided to us as in certain parametric limits. This falls under *Data Scaling.*

Once the preprocessing is carried out successfully, our next task constitutes generating a model that the user can train using the data already available.

The next step constitutes deciding on which model to take up for the kind of data available. Considering that our data are related to healthcare statistics, and taking forward the idea of predicting outcomes based on the available data, we are using the ***Apriori Algorithm*** to understand what kinds of associations are present between the data.

Our model suggests the inputs to be at least a few of the following: *age, gender, whether immunized at birth, smoking habits, filtered v/s unfiltered drinking water, any chronic diseases, blood pressure, anemia, access to private cooking area, access to private toilets, cleanliness of localities, personal doctors, cooking fuel used. And we link this with a single parameter termed as "Overall Health" in our model.*

Our aim constitutes deriving a model that would find associations between all of the aforementioned parameters (based on the surveys conducted).

AIM: To generate a model that fits perfectly for determining health-related outcomes using the Apriori algorithm in such a way that it helps formulate associations in data available to us, and link it with the overall health of the person.

HOW THIS IS CARRIED OUT: To generate a model that fits perfectly with our aim, we consider four major parameters: *Support, Confidence, Length, and Lift.*

Support: Parameter that allows us to judge the number of occurrences of a particular instance in comparison to the total number of instances. A support of 0.5 would mean that a particular item in a list has 50% occurrence in a total of 100%.

***Confidence**:* This shows the chance of occurrence of the second event considering the first event is taking place.

If for a set: {'A', 'B'}, the confidence is 0.29, this means that if event A occurs, there is a 29% chance that event B would also occur.

***Lift**:* This parameter is similar to the confidence, with the difference that it also takes into account the popularity *(or support)* of the resultant parameter.

***Length**:* This signifies the length of each rule set being generated. A set {'A', 'B'} has a length of 2.

If for a set: {'A', 'B'}, we wish to find out the lift, it would mean: The occurrence of event B considering event A, but also considering the popularity of event B.

A lift value greater than 1 for a set {'A', 'B'} suggests that event B is likely to happen if event A happens. A lift value less than 1 for the same set would suggest that event B is less likely to happen if event A happens.

Once the parameters such as these are established, the next aim is to set minimum threshold values for each.

Minimum Support can be set to a value that can be found out based on the number of items we have in our dataset. This would represent the minimum number of occurrences of each set of instances. For our model, after multiple iterations, we figure that the perfect *Minimum Support = 20% (0.2).*

Minimum Confidence can be set based on how we wish to know the relation between the predictor and the predictant events. After multiple iterations of *rules* being generated in our tests, we recommend the *Minimum Confidence = 30% (0.3).*

Minimum Length is set based on the number of lengths of rules we wish to generate. We wish our rules to generate at least 4 associations for the final outcomes. So, *Minimum Length = 4.*

Next, we generated *rules* significant to our model. These *rules* are generated and listed for the end user to find out associations between the data.

Here is an *example test rule* that we generated during the testing of the model:

```
v 62 = {RelationRecord} RelationRecord(items=frozenset({'no immune ', 'unhealthy ', 'smokes ', 'male ', 'unfiltered water'}), support=0.4,
   > 0 = {frozenset} frozenset({'no immune ', 'unhealthy ', 'smokes ', 'male ', 'unfiltered water'})
     1 = {float} 0.4
```

This rule takes in the following association:

{'no immune', 'unhealthy', 'smokes' 'male' 'unfiltered water'}

OUTCOMES FOR THIS RULE

(a) The support for this rule turned out to be a massive 0.4

(b) Out of the multiple Order Statistics derived from this rule, the one that has the maximum *lift* is: ***{'no immune', 'smokes', 'unfiltered water', 'male'} -> {'un-healthy'}***

The *lift* for this is 2.5, the confidence a straight 1.0 in our model.

This signifies in the model that a man who has not been immunized during his childhood, who smokes, and drinks unfiltered water has a good 100% of considering himself unhealthy.

This checked perfectly with our surveys conducted of people from across the country.

This signifies a model that has been generated with the exact *minimum confidence, length, support, and lift* to generate associations between the eating, drinking and smoking habits of people, their prior health conditions, and finally their overall health.

The model generation for health-related data suggested would include the following steps:

```
Collect Data
    Preprocess Data: Missing Values, Duplicates, Data Scaling, Data Slicing
        Generate Model:
            Use Apriori Algorithm
                Add Suggestive Minimum Support, Confidence, Lift, Length for Associations
                    Create Rules
                        Analyse Rules
                            Generate Result Lists from Rules
```

CONCLUSION

The study concludes that with the India becoming one of the world's fastest growing economies, there is still an extremely large part of the country lying under the poverty index, with more than 250 million citizens still marked as poor. Adding to that, statistics from the World Bank Organization suggest that more than 60 million children in the country are malnourished; with about 45% stunted (too short for their age), 20% wasted (too thin for their height, indicating acute malnutrition), 75% anemic, and 57% Vitamin A deficient. These statistics lead to an incriminating look into how the country's rural poor are reeling with health-care and nutrition needs. With technology advancing to greater heights in the field of data analytics, this paper aims at using big data analytics concepts such as data munging, visualization, clustering, and the likes into developing a single algorithm that would establish correlation between a person's nutrition and day-to-day activities with their health care patterns, to provide better healthcare predictions.

REFERENCES

1. Alkhatib, Ahmad Mohammad, Khloei-Talaei, Amir, Ghapanchi, Hossein Amir, 2015, 'Analysis of Research in Healthcare Data Analytics'
2. Bhattacherjee, A & Hikmet, N 2007, 'Physicians' resistance toward healthcare information technology: a theoretical model and empirical test', European Journal of Information Systems: Including a Special Section on Healthcare Information, vol. 16, pp. 725–37.
3. Conley, E., Owens, D., Luzio, S., Subramanian, M., Ali, A., Hardisty, A., & Rana, O. (2008), 'Simultaneous trend analysis for evaluating outcomes in patient-centred health monitoring services', Health Care Management Science, vol. 11, no. 2, pp. 152-66.
4. Jacob, S. 2012,'Young parkland physician makes a splash with predictive modeling software', D healthcare daily, Dallas.
5. Joumard, Isabelle & Kumar, Ankit, Organisation for Economic Co-operation and Development, 2015, 'Improving Health Outcomes and Health Care in India'
6. Russom, P. 2011, 'Executive summary big data analytics', Tdwi best practice report, Renton, Washington.
7. Swacchta Status Report, 2016, National Sample Survey Office

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Computer Application & Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you to supply your critical comments and suggestions about the material published in this issue, as well as on the journal as a whole, on our e-mail **infoijrcm@gmail.com** for further improvements in the interest of research.

If you have any queries, please feel free to contact us on our e-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward to an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

Sd/-
Co-ordinator

DISCLAIMER

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals