

Detection and Prediction of Frequent Diseases in India through Association Technique using Apriori Algorithm and Random Forest Regression

P. Aiswarya¹, M. Bhanu Sridhar², L. Kavitha³

^{1 2 3}Department of Computer Science and Engineering
Gayatri Vidya Parishad College of Engineering for Women

Abstract: - Data Mining is a process of analyzing huge data from different perspectives and summarizing it into useful information. The information can be converted into knowledge about historical patterns and future trends. Health care industry generates large amount of complex data about patients, hospitals resources, diseases, diagnosis methods and electronic patients' records. The data mining techniques are very handy to take appropriate medicinal decisions in curing diseases. The healthcare data can be "mined" to discover hidden information or patterns for effective decision making. The discovered knowledge can be utilized by the healthcare administrators to improve the quality of service and provide better facilities to the patients. Mining of frequent diseases helps clinicians to take better diagnosis decisions in curbing the occurrence of these diseases in the society. A systematic approach with the association rule based Apriori data mining technique is proposed in this paper to identify frequency of diseases. Further, prediction the frequent ones in a particular geographical area at a given time period has also been given a look through Random Forest Regression so as to be a boon to the lives of patients in remote areas.

Keywords: Data Mining, Apriori Algorithm, Random Forest Regression, prediction.

I. INTRODUCTION

Computer Science is a field which consists of various techniques and technologies related to Data Science, Artificial Intelligence, Internet of Things and Graphical Visualization. Data Mining can be described as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes [1]. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation.

Data Mining consists of four major functionalities which include Pre-processing, Association, Classification and Clustering. Association is a data mining function that

discovers the probability of the co-occurrence of items and interesting patterns in a collection. Association is a process of generating frequent item sets based on few thresholds. The relationships between co-occurring items are expressed as association rules. The association technique works based on three metrics, they are support, confidence and lift. Attributes such as the state name, disease, number of cases in a particular time period are used to find the frequent diseases in a particular geographical area.

This paper proposes a methodology that utilizes Association technique to generate interesting patterns from the health data. We use the Apriori Algorithm to mine the frequent diseases across each state in India based on the minimum support threshold. The paper is organized as follows: Section I of the paper deals with introduction; Importance of health data is described in Section II, in Section III categories of diseases in India are presented; Apriori algorithm and Random Forest Regression is presented in Section IV; Section V deals with the methodology together with experimental results and finally the conclusion is presented in Section VI.

II. IMPORTANCE OF HEALTH DATA

Health data is considered as one of the most dynamically generated data [2]. A gradual increase in the number of health cases registering across the world every day can apparently be observed. This data is considered as most gripping and challenging dataset collection to be mined and is used to obtain interesting and hidden patterns. Mining of health data and prediction of future condition not only helps the clinicians and medicos to improve their quality of service in treating the patients but also in determining the futuristic ailments.

Knowledge discovery from health data helps to identify frequent diseases in a particular geographical area at a given time period to provide a better understanding of the root cause of the complaint. This gives a clear picture to the researchers and helps them to identify different aspects triggering the particular disease, so that they can uproot it from the origin. If this is done, proper vaccines and medicines can be used and awareness can be spread among public to take necessary precautions to prevent the diseases

so that the epidemics/pandemics can be efficiently controlled. This surely reduces the mortality rate and helps the society to lead a healthy life.

India has witnessed remarkable progress in the health status of its population. However, over the past few decades, there have been major transitions in the health care field that had serious impact on health of patients. Changes have come up in economic development, nutritional status, fertility and mortality rates. Consequently, the disease profile has changed considerably. Communicable diseases such as Malaria, Kala-azar, Dengue, Chikungunya and Acute-Encephalitis have been recorded as the most frequently occurring diseases in the country. Though there have been substantial achievements in controlling communicable diseases, they still contribute significantly to disease burden of the country. Decline in mortality from communicable diseases have been accompanied by a gradual shift to the prevalence of chronic non-communicable diseases (NCDs) such as cardiovascular disease (CVD), diabetes, chronic obstructive pulmonary disease (COPD), cancers, mental health disorders etc.

III. DISEASES PREVALENT IN INDIA

India is the second most populous country in the world and with almost one-fifth of the world's population living in India, the health status and the drivers of health loss is expected to vary between different parts of the country and between the states.

Accordingly, effective efforts to improve population health in each state require systematic knowledge of the local health status and trends. The diseases which tend to occur in the states can be broadly classified into two types: Communicable and Non-Communicable diseases.

A. Communicable Diseases

The authentic data by The National Health Profile 2016, 2017, 2018 and 2019 reports the following diseases:

• MALARIA

Malaria is a mosquito-borne infectious and communicable disease with symptoms which typically include fever, tiredness, vomiting, and headache. This disease is spread by single-celled microorganisms of the Plasmodium group and is most commonly spread by female Anopheles mosquito [3].

• KALA-AZAR

Visceral leishmaniasis (VL), also known as kala-azar, is the most severe form of leishmaniasis and, without proper diagnosis and treatment, is associated with high fatality. This disease is caused by protozoan parasites of the genus Leishmania.

• CHIKUNGUNYA

Chikungunya is an infection caused by the chikungunya virus (CHIKV). Symptoms include fever and joint pains. Other symptoms may include headache, muscle pain, joint

swelling, and a rash. The disease is caused by two mosquito breeds namely Aedes albopictus and Aedes aegypti [4].

• DENGUE

Dengue fever is caused by the dengue virus. The symptoms may include a high fever, headache, vomiting, muscle pains, joint pains, and a skin rash. Dengue is spread by several species of female Aedes mosquito, mainly A. aegypti.

• TYPHOID

Typhoid fever is a serious disease spread by contaminated food and water. Symptoms of typhoid include high fever, weakness, stomach pains, headache, and loss of appetite. This is a bacterial infection spread due to Salmonella.

• PNEUMONIA

Pneumonia is a lung inflammation causing high respiratory problems such as cough, fever, chills and difficulty in breathing.

• ENCEPHALITIS

Encephalitis is an inflammation of the brain which is mostly caused by Viruses. Various viruses causing encephalitis include herpes viruses, West Nile, Japanese encephalitis, and tick-borne viruses.

• JAPANESE ENCEPHALITIS

This disease is caused by Japanese encephalitis virus and is spread by Culex mosquitoes and they bite mainly during the night or just after sunset.

The following pie-charts depict the morbidity rates from the years 2017 and 2018 [5].

Percentage distribution of Morbidity Reported in Communicable Diseases - 2017

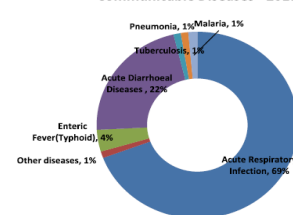


Figure 1: Morbidity rates in 2017

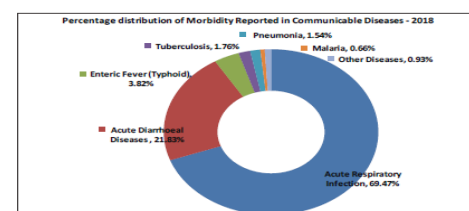


Figure 2: Morbidity Rates in 2018

B. Non-Communicable Diseases

Non-communicable diseases (NCDs) encompass a vast group of diseases such as cardiovascular diseases, cancer, diabetes and chronic respiratory diseases. NCDs contribute

to around 5.87 million (60%) of all deaths in India and to about 82% of deaths in the world [6]. Four NCDs mainly responsible for the mortality and morbidity include the following:

- **CARDIOVASCULAR DISEASES**

Cardiovascular disease (CVD) involves the heart or blood vessels. Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, vascular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous-thrombosis [7].

- **CANCER**

Cancer, also called malignancy, is an abnormal growth of cells. There are more than 100 types of cancer, including breast cancer, skin cancer, lung cancer, colon cancer, prostate cancer, and lymphoma. Symptoms vary depending on the type [8].

- **CHRONIC RESPIRATORY DISEASE**

Chronic respiratory diseases are chronic diseases of the airways and other parts of the lung. Some of the most common are asthma, chronic obstructive pulmonary disease (COPD), lung cancer, cystic fibrosis, sleep apnoea and occupational lung diseases [9].

- **DIABETES**

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. This damages especially the nerves and blood vessels.

IV. APRIORI ALGORITHM AND RANDOM FOREST REGRESSION

A. Apriori Algorithm

Apriori algorithm finds its place in the association functionality of data mining where this algorithm is especially used for mining frequent item sets from a large dataset. The major advantage of this algorithm over other association algorithms is that, Apriori algorithm can work on large datasets and also is quite easy to understand and implement. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. This algorithm is being applied in domains such as Market Basket Analysis and Hospital Information Systems.

Apriori algorithm works on the Apriori property which says

- An item set is frequent when its subsets are frequent
- An item set to be infrequent if its supersets are infrequent

So, this becomes a great way of pruning the dataset in extracting the frequent trends in the same. The Apriori algorithm does the two following operations:

- **Join Operation:** Candidate itemset is generated by joining the previous frequent itemsets with itself.

- **Prune Operation:** Any itemset that is not frequent cannot have a subset which is frequent (or) any itemset that is not frequent cannot be a super set of a frequent itemset.

Apriori algorithm is the most classical and important algorithm for mining frequent item sets. Based on the Apriori principle any subset of a frequent itemset must also be frequent. Example: if {XY} is a frequent itemset, both {A} and {B} must be frequent item sets [10]. The main idea of Apriori algorithm is to make several passes over the database. It employs an iterative approach known as a breadth-first search. In each subsequent pass, we begin to add the items into an itemset, called as candidate itemset. At the end, we determine which of the candidate item sets are actually large or frequent, until no more frequent itemsets can be found.

Hence the above theory suggests that Apriori follows bottom-up approach in scanning and pruning the dataset.

The steps in applying the Apriori algorithm to a dataset are given below:

Step 1: Consider a dataset consisting of 'n' number of transactions.

Step 2: Now calculate the minimum support count based on the transaction count.

Step 3: When the algorithm is applied on the dataset, the items whose support count is greater than or equal to the minimum support threshold are added into the candidate item set in each of the iteration until the candidate set gets an empty value, that is, ' ϕ ' value.

Step 4: Having found all the frequent item-sets, the algorithm gets terminated at this point.

The pseudo code for this algorithm can be given as below:

C_k : Candidate itemset of size k

L_k : frequent itemset of size k

$L_I = \{\text{frequent items}\};$

for ($k = 1; L_k \neq \emptyset; k++$) **do begin**

$C_{k+1} =$ candidates generated from L_k ;

for each transaction t in database **do**

increment the count of all candidates in C_{k+1} that are contained in t

$L_{k+1} =$ candidates in C_{k+1} with min-support

end

return $\cup_k L_k$;

- **Random Forest Regression**

Random Forest Regression uses ensemble learning method for regression and to find correct predictions without over-fitting and maintaining the relationship between the dependent and independent variables. Ensemble learning method is used to combine multiple machine learning algorithms to make accurate predictions when compared to an individual model. Random Forest Regression follows the Bootstrap Aggregation or bagging kind of ensemble technique to make each model in the training set run independently and then compilation of the outputs is done at the end. Random Forest Regressor operates by constructing various decision trees at training time and does the mean prediction of the individual trees. The primary advantage of Random Forest Regression over

other regression techniques is that, Random Forest Regressor seldom overfit any model and can also work efficiently on large and dynamic datasets like the Hospital and Health datasets.

V. METHODOLOGY AND RESULTS

The orientation provided by our work can be very competent in finding the frequent diseases in a particular geographical area and predicting the trend of the cases in the future. This leads to improved decision-making and treatment as India has ample evidences of impacts attributed to mismatch between disease burden and its casual factors. Thus, interventions adopted for treatment and priorities in resource allocation can also be put into their appropriate places.

This paper deals with the demography of India, which is associated with a great fluctuation in anamnesis of the nation with a vision of predicting the frequently occurring diseases at various locations.

A. The Dataset

The authentic instances of the dataset have been derived from a series of editions published as the National Health Profile by the Central Bureau of Health Intelligence, Government of India, along with the collaboration of World Health Organization (WHO), various Central Ministries and all the state/union territory health departments. The data depicted in these editions provide

comprehensive information related to health sector. These issues of the National Health Profile from 2015-2019 provide vital information on all major health sector related indicators, demography, socio-economics, health status, health finance, health infrastructure and human resources for the specified calendar year. The dataset consists of seven attributes:

- State
- Disease
- Number of cases_2015
- Number of cases_2016
- Number of cases_2017
- Number of cases_2018
- Total number of cases

Each tuple shows the trend of the number of cases reported of a particular disease.

B. Proposed Work

An updated and reliable health database is the foundation of decision-making across all health system building blocks. This is essential for health system policy development and implementation, governance and regulation, health research, human resources development, health education/training, service delivery and financing. To achieve some of these objectives, Central Bureau of Health Intelligence collects data from various sectors, ensuring their overall quality, relevance and timeliness.

1	State	Disease	Number of cases_2015	Number of cases_2016	Number of cases_2017	Number of cases_2018	Total Number of Cases
2	AndhraPradesh	Malaria	25042	23613	16972	6034	71661
3	AndhraPradesh	Chikungunya	817	960	1162	622	3561
4	AndhraPradesh	Kala-azar	0	0	0	0	0
5	AndhraPradesh	Acute-Encephalitis-Syndrome	50	4	37	16	107
6	AndhraPradesh	Japanese-Encephalitis	0	0	1	0	1
7	AndhraPradesh	Dengue	3159	3417	4925	4011	15512
8	AndhraPradesh	Cholera	0	0	0	0	0
9	AndhraPradesh	Acute-Diarrhoeal-Diseases	1122740	1194005	1189407	1193473	4699625
10	AndhraPradesh	Typhoid	146385	170249	163929	148490	629053
11	AndhraPradesh	Acute-Respiratory-Infection	2311538	3049810	3103021	3149897	11614266
12	AndhraPradesh	Diphtheria	57	0	6	67	130
13	AndhraPradesh	Whooping-Cough	83	11	1	187	282
14	AndhraPradesh	Measles	44	115	65	184	408
15	AndhraPradesh	Viral-Hepatitis	3358	2402	7382	8817	21959
16	AndhraPradesh	Pneumonia	24126	24745	32636	37749	119256
17	AndhraPradesh	Meningococcal-Meningitis	86	30	552	758	1426
18	AndhraPradesh	Syphilis	3948	2413	5395	4069	15825
19	AndhraPradesh	Gonococcal-Infection	19021	17131	13996	12484	62632
20	AndhraPradesh	Tuberculosis	28714	41790	41106	48070	159680
21	AndhraPradesh	Swineflu	258	12	476	402	1148
22	AndhraPradesh	Chickenpox	257	578	1484	812	3131
23	AndhraPradesh	Encephalitis	7	3	638	768	1416
24	AndhraPradesh	Viral Meningitis	9	29	1493	684	2215
25	AndhraPradesh	Diabetes	49620	58608	271912	391843	771983
26	AndhraPradesh	Hyper tension	51831	60055	270812	399094	781792
27	AndhraPradesh	Cardiovascular	1881	5288	8437	9585	25191
28	AndhraPradesh	Cancer	232	2818	803	5705	9558
29	Arunachal Pradesh	Malaria	5088	3128	1546	625	10387
30	Arunachal Pradesh	Chikungunya	35	239	133	507	914
31	Arunachal Pradesh	Kala-azar	0	0	0	0	0
32	Arunachal Pradesh	Acute-Encephalitis-Syndrome	73	1	5	0	79
33	Arunachal Pradesh	Japanese-Encephalitis	32	0	5	0	37
34	Arunachal Pradesh	Dengue	1933	13	18	1	1965
35	Arunachal Pradesh	Cholera	0	9	0	0	9
36	Arunachal Pradesh	Acute-Diarrhoeal-Diseases	10834	12746	23888	22972	70440
37	Arunachal Pradesh	Typhoid	4476	4582	3553	2458	15069
38	Arunachal Pradesh	Acute-Respiratory-Infection	15840	21670	36250	33668	107428
39	Arunachal Pradesh	Diphtheria	283	0	0	0	283
40	Arunachal Pradesh	Whooping-Cough	24	49	15	34	122
41	Arunachal Pradesh	Measles	17	365	318	35	735
42	Arunachal Pradesh	Viral-Hepatitis	292	273	140	70	775
43	Arunachal Pradesh	Pneumonia	577	484	1181	707	2949
44	Arunachal Pradesh	Meningococcal-Meningitis	0	4	0	0	4
45	Arunachal Pradesh	Syphilis	22	36	16	2	76
46	Arunachal Pradesh	Gonococcal-Infection	2	3	8	2	15
47	Arunachal Pradesh	Tuberculosis	837	1117	1332	1780	5066
48	Arunachal Pradesh	Swineflu	0	0	5	0	5
49	Arunachal Pradesh	ChickenPox	158	578	400	309	1445
50	Arunachal Pradesh	Encephalitis	0	0	2	0	2

Figure 3: Sample Dataset

Data is then converted into relevant information to support planning, management, and decision making.

There are totally 27 diseases in the dataset whose frequency across India's 29 states and 7 union territories have been deducted.

The methodology of our proposed work follows the steps below:

Step1: First, the cumulative number of cases for each state and union territory has been calculated based on the total number of cases attribute from the dataset.

Step2: This cumulative number of diseases can be considered as the total number of transactions registered for a particular state (or) union territory (UT).

Step3: Now the minimum support threshold is set. Since the dataset is quite large in this case, the minimum support threshold is taken within the range of 1% to 5% of the total number of transactions. Formula for the same is taken as:

Min. Support (State/UT) = $(0.01/0.02/0.03/0.04/0.05) \times \text{total_transactions of State/UT}$

Step4: All the states and union territories need not have the same health indicators such as demography, climatic conditions, health finance infrastructure; accordingly, the occurrence of the diseases may vary from state to state. That is the reason why a unique minimum support threshold is considered for each state and union territory so

as to identify the epidemics that are recurrent to that particular place.

Step5: After the minimum support threshold is set, the diseases whose total number of cases are greater than or equal to the minimum support threshold can be added into the candidate set as a frequent disease. This process is done likewise for all the other 26 diseases to find out the frequent diseases in that particular

State/UT and with the help of user-interactive window, the user can select the state whose frequent disease he/she would like to view along with the for the same. Another window of the same purpose is used to display the graphs of diseases which show the areas of India in which that disease has been reported.

```

def aprialgToFindFrequentDiseases(data):
    dataset=loaddataset(data)
    print("enter the minimum support:")
    min1=float(input(text1.value))
    minsup={}
    sum=0
    temp=0
    n=len(dataset)
    for i in range(1,n-1):
        if(dataset[i][0] in minsup):
            sum+=int(dataset[i][6])
            temp=1
        else:
            if(temp==1):
                minsup[dataset[i-1][0]]=int(sum*min1)
            sum=int(dataset[i][6])
            temp=0
            minsup[dataset[i][0]]=0
        minsup[dataset[i-1][0]]=int(sum*min1)
    return(minsup,n,dataset)

minsup,n,dataset=aprialgToFindFrequentDiseases('Dataset.csv')
print(minsup)

enter the minimum support:
0.01
{'AndhraPradesh': 190118, 'Arunachal Pradesh': 3052, 'Assam': 12990, 'Bihar': 70282, 'Chhattisgarh': 36926, 'Goa': 6181, 'Gujarat': 121725, 'Haryana': 157055, 'Himachal Pradesh': 83388, 'J&K': 59504, 'Jharkhand': 23159, 'Karnataka': 140355, 'Kerala': 259254, 'Madhya Pradesh': 108901, 'Maharashtra': 125403, 'Manipur': 4394, 'Meghalaya': 23873, 'Mizoram': 3739, 'Nagaland': 2887, 'Odisha': 145501, 'Punjab': 46072, 'Rajasthan': 231769, 'Sikkim': 6778, 'Tamil Nadu': 134809, 'Telangana': 79452, 'Tripura': 12658, 'Uttarakhand': 18311, 'Uttar Pradesh': 198794, 'West Bengal': 240150, 'Andaman&Nicobar Islands': 23012, 'Chandigarh': 9982, 'Dadra&Nagar Haveli': 5314, 'Daman&Diu': 3913, 'Delhi': 22626, 'Lakshadweep': 1781, 'Puducherry': 42358}

```

Figure 5.1. Results after finding the minimum support threshold for every state and union territory

In Andhra Pradesh frequently occurring diseases are ==> Acute-Diarrhoeal-Diseases, Typhoid, Acute-Respiratory-Infection, Diabetes, Hyper tension

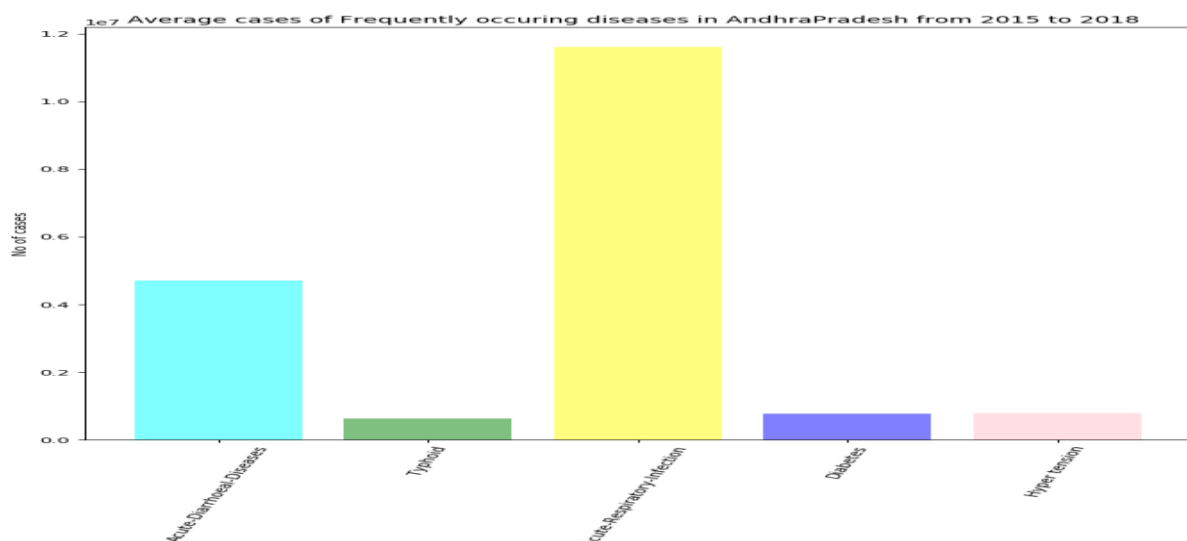


Figure 5.2: State-wise Frequent Diseases and Graph

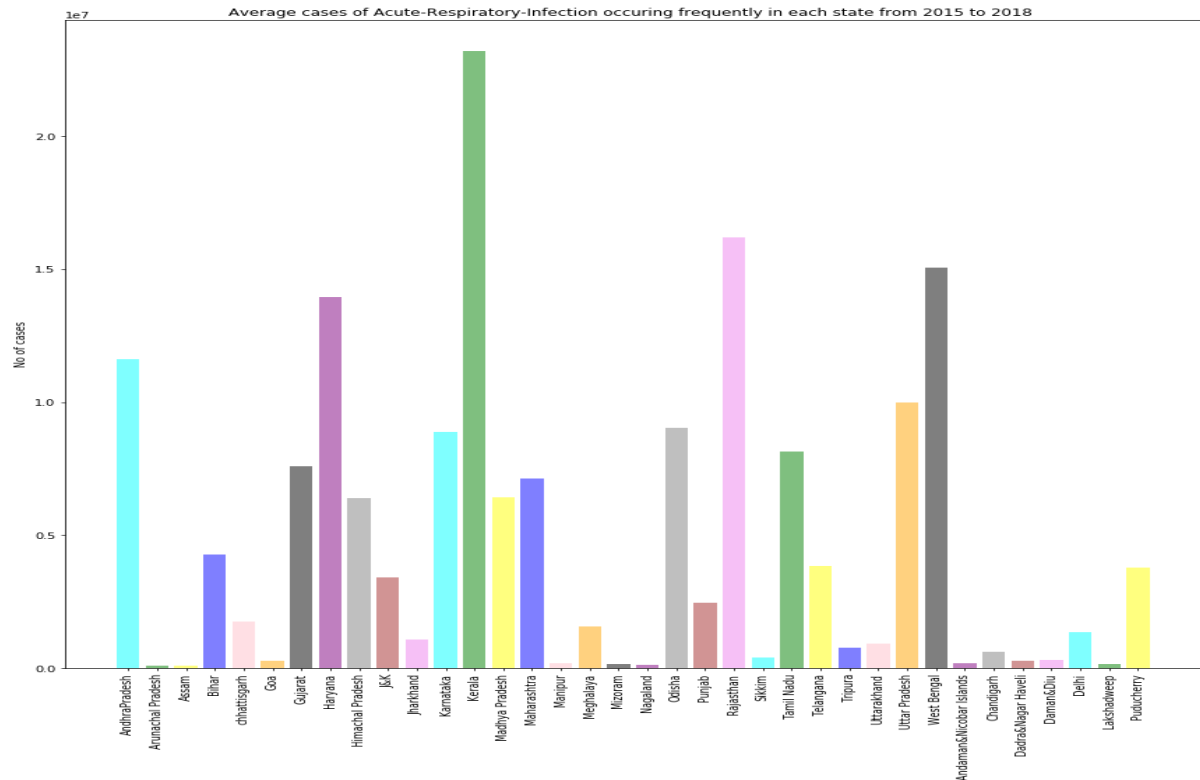


Figure 5.3: Plot for Acute-Respiratory Infection Prevalence in India

Step6: The purpose of finding the frequent diseases is served by predicting the number of cases that would occur in the future based on the previous trends or values with the help of Random Forest Regressor. Random Forest Regressor has been implemented specifically for our work

as this regressor does not overfit the training samples thus improving the accuracy or the r^2 score which is used to measure how close the data is fitted or predicted. This can be quoted as an extension to the concerned work.

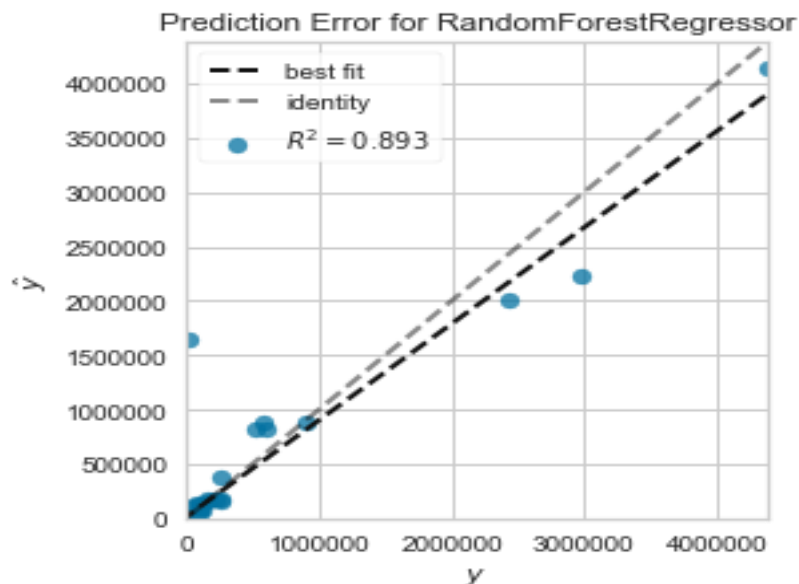


Figure 6.1: The Prediction Error Graph plotted for the X and y test values

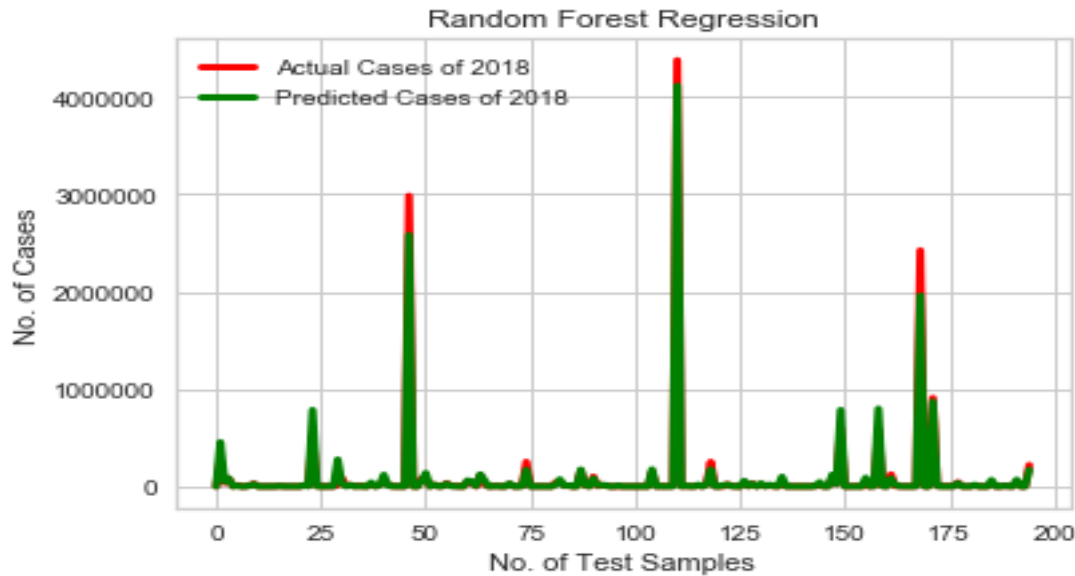


Figure no. 6.2 Plot for the Actual and Predicted values (through Random Forest)

Step7: The diseases that are infrequent and the diseases that are frequent at a very few areas of India are also listed out. This helps to identify diseases whose rate of

occurrence is quite less when compared to others and can be curbed or controlled in the very budding stage itself.

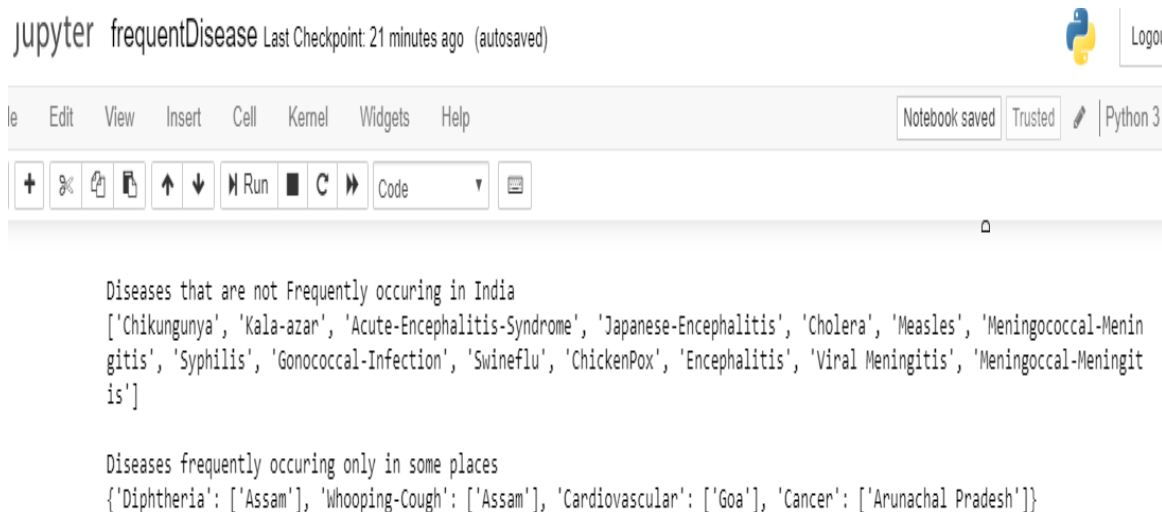


Figure 7: Result showing the list of diseases that are less frequent

V.III. FUTURE SCOPE

The proposed method is useful to identify the frequent diseases in a large medical dataset collected from an authentic source in India during a stipulated time period from 2015-2018. The end result of this research will help the clinicians and other decision-making forces in making medical decisions for frequently occurring diseases. The predictions and analysis made can be furthered to classify the diseases into three to four classes based on the rate of occurrence of the diseases and take appropriate decisions as to control the problems faced by the patients and the society. Still yet, this work can be extended to work on unexpected pandemics and predictions based upon which a country can be ready to face the invisible enemy.

V. CONCLUSION

In this paper, a methodology has been presented for analysing the health data of diseases in different areas or states of India. The dataset taken up for consideration is examined and considered for utilization as per the requirements. By using association principle through Apriori algorithm through the steps described, the frequent diseases in different areas has been obtained and presented in different angles.

The ultimate idea behind all this process is after discovering the frequent diseases in different states or UTs in India, necessary steps to be taken for their prevention in the coming years and keeping the concerned medicines ready for them before the next year or season. This apparent prediction can be a great help for the local

medicos and most importantly, for the new patients in remote areas. This can be a great help for the society and protect the patients in remote areas by offering the exactly required medicines, thus prevent deaths. This problem recurs every year; by utilizing this idea, it can be tackled in the best way possible. Finally, the real idea of the work is to protect the people as best as possible and help the local governing bodies and hence, the society.

REFERENCES

- [1] Ilayaraja M and Thiru Meyyappan, "Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm", IEEE Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, 2013.
- [2] American Journal of Public Health: Editorial - Addressing Disparities in the Health of American Indian and Alaska Native People: The Importance of Improved Public Health Data, Supplement 3, 2014, Vol. 104, No. 53.
- [3] Wikipedia: <https://en.wikipedia.org/wiki/Malaria>
- [4] Centres for diseases control and prevention, <https://www.cdc.gov/chikunguniya/index.html>
- [5] <https://www.encephalitis.info>
- [6] www.cbhidghs.inic.in
- [7] M. Bhanu Sridhar, Y. Srinivas, and M. H. M. Krishna Prasad, "Software Reuse in Cardiology Related Medical Database Using K-Means Clustering Technique", Journal of Software Engineering and Applications, 2012, 5, 682-686.
- [8] <https://www.webmd.com/cancer/default.htm>
- [9] <https://www.canada.ca/en/public-health/services/chronic-diseases/chronic-respiratory-diseases.html>.
- [10] Sanjeev Rao and Priyanka Gupta, "Implementing Improved Algorithm over APRIORI Data Mining Association Rule Algorithm", International Journal of Computer Science and Technology IJCST, Volume 3, Issue1, Jan. - March 2012, ISSN: 0976-8491.