Influenza Prediction

Akshay George Koshy,Yuvaana Sundarakrishnan,K. P. Vijayakumar

Abstract*: The point of the undertaking is to foresee the spread of a plague by breaking down the conditions in the territories where individuals are influenced. The momentum focal point of the task is one specific pestilence ailment, Influenza, which is an irresistible illness brought about by introduction to the flu infection. The expectation will be finished by breaking down the spread dependent on the development of the infection through the populace. To achieve such a forecast, an administered learning model was utilized on the informational collection assembled. The calculation utilized is Random Forest Algorithm (RFA), which is basic and is typically utilized for arrangement and relapse.

Keywords* : *Random forest algorithm (RFA), Prediction model, Influenza,

I. INTRODUCTION

Flu is a typical irresistible sickness cause by the flu infection. Side effects can change from mellow to genuine. Signs consolidate fever, runny nose, sore throat, muscle agony, hacking and sneezing. The hack may persevere for about fourteen days from beginning of the disease. Three of the four sorts of influenza contaminants impact individuals: Type A, Type B, and Type C. Type D has not been known to taint individuals, anyway is acknowledged to can do all things considered. The disease is spread through the air and is acknowledged to spread simply over short illnesses. Testing the throat, sputum, or nose can assert the end

TYPE A

Aquatic birds are the common hosts to this type of influenza. The infections it causes can adjust and affect different species of poultry and people, which can lead to the flare of a pandemic. The infections caused by this type is the most harmful (to humans) among the four different pathogens.

TYPE B

This type of flu contaminates only humans and is less likely to be contracted than fly type A. This type of flu is less hereditarily diverse as it transforms at a rate of 2 to 3 times slower than type A. Hence, a degree of immunity to this flu can be procured at an early age. Nevertheless, an enduring resistance is not possible. The seal and ferret are known to be powerless to this type.

TYPE C

This type of infection contaminates dogs, pigs and people can cause extreme illnesses and localized epidemics. This type is, however, less common than the other sorts and usually causes mild illnesses in children.

TYPE D

The type D of influenza affects pigs and cattle. There is a chance that it could also affect people, but so far no such cases have been observed. Currently, this type is not associated with any major epidemics.

II. EFFECTS OF ENVIRONMENTAL FACTORS

Influenza virus is heavily affected by temperature and precipitation, it is found with increase in temperature the virus is unable to spread as quickly, bring inefficient at 30C. Higher temperature facilitate in the increase of the range of disease vectors. Dry conditions are also found to supplement the spread of the virus, then when it more humid.

Most by far of forecast calculations checked on so far actualize Support Vector Machine (SVM), Social media, Artificial Neural Networks (ANN, etc to envision the flare-up/spread of a malady. The frameworks that are express to a particular ailment are more precise in their expectations than frameworks that are not unequivocal to a particular sickness. This has something do with the informational index that it uses to make said expectations. The greater part of the flow frameworks are utilized to anticipate episode as opposed to foresee the spread of the malady through the populace.

Certain ecological states of a particular domain were taken as the informational indexes and using that testing data, forecasts were made with respect to whether the region will be influenced. This methodology would be advantageous for the proposed model as it is has a comparable diagram, then again, actually since they are various maladies, they would require distinctive testing information to make forecasts. Sharma et al. (2015) [1].

Making forecasts utilizing content investigation, supposition examination and groupings utilizing a directed learning model on information gathered. Long range informal communication stages was utilized as the information source and information was

Revised Manuscript Received on September 22, 2019.

Akshay George Koshy,Computer Science and Engineering,SRM Institute of Science and Technology,akshay.g.koshy@gmail.com
Yuvaana Sundarakrishnan,SRM Institute of Science and Technology,yuvaana.sk@gmail.com
K.P.Vijayakumar,SRM Institute of Science and Technology,vijayakumar.kp@ktr.srmuniv.ac.in

recovered from constant sources. Utilization of notion investigation means doing expectations utilizing the seriousness of the pestilence dependent on the internet based life spread.

Ranjan and Misra (2014) [2]. Use of Spatial and Covariate Information (SCI) from dispersed sources to improve the effectiveness of flare-up recognition. A repressed methodology instead of others as there

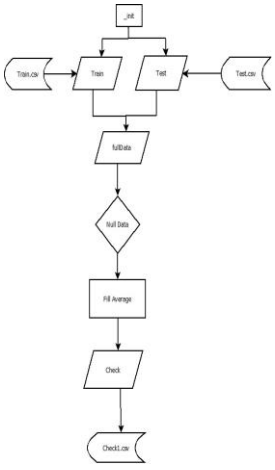
were constraints related with utilizing huge volumes of information to play out the expectations. Consequently it was found that the information size must be held under tight restraints to approach the greatest proficiency of the calculation. Buckeridge et al. (2005) [3].

Multivariate Hidden Markov Model (MHMM) was utilized to recognize the spread of flu utilizing information assembled from heterogeneous information gathered from sources, for example, Google inquiry information to foresee flare-ups. It stressed the need to likewise consider geological condition when thinking about the spread of the ailment. Zhou et al. (2018) [4]. Ayiad and Fatta (2017) [6].

Distinguishes associations of a tainted hub utilizing cell phone based individual and network detecting to check the cooperations of the contaminated hub with different hubs and along these lines foreseeing the wellbeing of the hub. This model predicts the malady as for a solitary hub by anticipating the contaminated hub. Ranjan and Misra (2014) [7].

III. SYSTEM DESIGN

Upon initialization, the data is split into training and testing data respectively. The training data and testing data are initialized separately. Train is the variable holding training data and Test is the variable holding the testing data. After training and testing data are processed separately, the data is combined, giving rise to fulldata. The fulldata is checked for any null data spots. The null data spots (if any) are filled with average values to make the data complete. After filling the null spots and obtaining the completed data set, the data is checked, and the whole data is processed



(a) Architecture Diagram.

IV. WORKING

DATA SET

It is proposed to use data gathered from FluNet, World health organization-based influenza surveillance system, to load the number of cases over a period of 10 years from 2009-2018. Additional factors like temperature (average temperature of the country over a month), precipitation (average precipitation in the country monthly) and population are also taken into account. Temperature and precipitation were taken from the world bank database. 4.1 The environment used is anaconda with jupyter notebook which eases the initialization of required packages used for the program. The data obtained from FluNet in conjunction with the World Bank data will be introduced as a .csv file which will be used to train the model. The data is run through an Machine Learning (ML) algorithm Random Forest Algorithm (RFA) which will then train the model and when test.csv (the data to be tested for) is introduced the model will output the ID of the specific set of data with the probability of such an occurrence

MODEL

The environment used is anaconda with jupyter notebook which eases the initialization of required packages used for the program. 4.2 When test.csv (the data to be tested for) is introduced the model will output the ID of the specific set of data with the probability of such an occurrence. 4.2 The program is trained using RFA which is a simple easy to use algorithm, which provides an accurate output. 4.2 RFA works by building ensemble of decision trees, trained most of the time with bagging method. This works by combining learning models to increase the overall result. RFA also adds randomness to the model thus increasing the diversity of results.

ALGORITHM

RFA is an algorithm capable of both classification and regression. As with most cases the greater number of trees in a model proportionally increases the accuracy of the model. i.e the higher the number of trees the better the accuracy of the model

Reasons for using RFA:

1. RFA can handle missing values.
2. The greater number of trees RFA won' t overfit the model.
3. Can use categorical values with RFA.

RFA pseudocode:

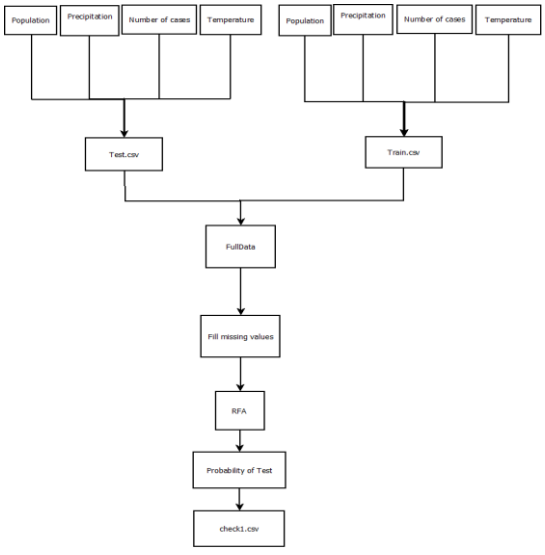
1. Select a random number of k feature from the total x features x should be
greater than k.
2. With respect to k feature(s), formulate the node z using best split point.

ID	Cases	Temperature	Precipitation	Population
361	361	-23.226	27.4503	33.7
361	78	-20.737	25.7178	34.1
363	6049	-21.626	28.0723	34.5
364	546	-22.644	27.8012	34.9
365	13929	-23.274	24.7111	35.2
366	11457	-21.165	32.0206	35.6
367	13032	-22.462	30.0385	35.9
368	2267			36.2
369	11930			36.6
370	21996			36.9
371	3601	-22.483	20.4024	33.7
372	31	-17.658	17.3839	34.1
373	4896	-22.255	22.0558	34.5
374	2512	-18.766	20.1216	34.9
375	3806	-21.793	18.8119	35.2
376	4912	-23.899	19.879	35.6
377	6261	-24.541	23.0041	35.9

3. Split into more child nodes using the best split.
4. Repeat step 1 to 3 till you get a set number of nodes.
5. Build the forest using step 1 to 4 n times to get n number of trees.

RFA prediction pseudo code :

1. Test features and rules of each randomly created tree is used to predict a set of output, which is then stored.
2. Formulate votes for each target output.
3. The target with highest votes is taken as final prediction



(a)Working of model

ID	Cases	Temperature	Precipitation	Population
1	38072	-2.0841	37.5391	306.77
2	35526	-0.7792	50.1073	309.34
3	44803	-5.6359	37.7097	311.64
4	21868	-5.4456	45.2091	313.99
5	74599	-4.1921	51.3893	316.23
6	78425	-3.6393	35.9334	318.62
7	181857	-4.1873	51.3893	321.04
8	85995		41.7446	323.41
9	162355			325.72
10	305390			329.1
11	48250	2.14271	35.2346	306.77
12	33324	-2.461	39.7066	309.34
13	59407	-4.2505	47.9114	311.64
14	25920	-1.4537	47.9114	313.99
15	59419	-3.3945	42.0856	316.23
16	42863	-4.6473	43.3208	318.62
17	103149	-3.4854	44.1704	321.04
18	119479		39.1029	323.41
19	208511			325.72
20	321339			329.1

(b) Input table (Training)

V. ADVANTAGES

- Probability of a outcome can be accurately predicted.
- Easy to update when required.
- Easy to adapt for other diseases.
- Gives ample time to prepare for worst case scenario.

VI. DISADVANTAGES

- Requires large set of data.
- Focuses on country and does not take into account variation across the country.
- Can only accurately detect one disease per set due to variation in disease
- infection parameters.

(c) Input table (Testing)

	ID	Probability of Cases
40	401	0.002
41	402	0.003
42	403	0.005
43	404	0.005
44	405	0.001
45	406	0.015
46	407	0.001
47	408	0
48	409	0
49	410	0
50	411	0.008
51	412	0.004
52	413	0.004
53	414	0.022
54	415	0.044
55	416	0.023
56	417	0.066

(d) Output

B. Abbreviations and Acronyms

- ANN - Artificial Neural Networks
- CNN - Convolutional Neural Network
- MHMM - Multivariate Hidden Markov Model
- RFA - Random Forest Algorithm
- SCI - Spatial and Covariate Information
- SVM - Support Vector Machine

VII. RESULT AND CONCLUSION

Influenza is an infectious disease that infects and kills thousands of people every year. It is of utmost importance that people be prepared for possible outbreaks. Influenza can easily be treated in most cases but to its ability spread quickly it can overwhelm the under-prepared. This project will allow us to predict the number of cases based off data that is readily available such as population, temperature and rainfall. The model also allows compensation for minor gap in data by substituting values as needed. The ability to predict the amount of cases for a specific set of factors allows us to be prepared for when an outbreak does happen and prevent unnecessary fatalities. This model can also serve as a basis for future models to build upon thus reducing the number of fatalities to a almost negligible level and saving millions from unnecessary harm.

VIII. ACKNOWLEDGMENT

We would very much like to thank Dr. K.P. Vijayakumar for his support and guidance in reviewing this research work.

REFERENCES

1) "Cloud task scheduling basedon load balancing ant colony optimization,"by K. Li, G. Xu, G. Zhao, Y. Dong, and D. Wang, Sixth Annu. Chinagrid Conf, pp. 3–9, Aug. 2011.

2) "Load Balancing of Nodes in Cloud Using AntColony Optimization". In proc. 14th International Conference on Computer Modeling and Simulation (UKSim), IEEE, pp:3-8, March 2012.

3) " Load Balancing The Essential Factor In Cloud Computing",by Mr. Jayant Adhikari, Prof. Sulabha Patil, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue. 10, pp. 1-5, 2012.

4) "Efficient and Enhanced Algorithm in Cloud Computing" International Journal of Soft Computing and Engineering (IJSCE)ISSN:2231-2307, Volume-3, Issue-1, Mfarch 2013.

5) "Dynamic load balancing: improve efficiency in cloud computing,"by A.Roy International Journalof Emerging Research in Management &Technology, vol. 9359, no. 4,pp. 78–82, 2013.

6) "Load balancing in cloud computing,"by R. Kaur and P. Luthra, Int. Conf. 2007. on Recent Trends in Information, Telecommunication and Computing, ITC, pp. 1–8, 2014.

7) "Load Balancing Algorithms in Cloud Computing: A Survey of Modern Techniques"by Sidra Aslam Munam, Ali Shah 2015 National Software Engineering Conference (NSEC2015),IEEE, pp:30-35

8) "Static Load Balancing Algorithms In Cloud Computing:Challenges & Solutions" International Journal Of Scientific & Technology Research Volume 4, Issue 10, October 2015

9) "A Comparative Study of Different Static and Dynamic Load Balancing Algorithm in Cloud Computing with Special Emphasis on Time Factor",by Abhijit Aditya, Uddalak Chatterjee and Snehasis Gupta, International Journal of Current Engineering and Technology, Vol.5, No.3 (June 2015).