Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.9 No.1, 2020, pp.14-18

DOI: https://doi.org/10.51983/ajcst-2020.9.1.2155

Influenza Prediction: Analyzing Machine Learning Algorithms

Sapna Yadav[1] and Pankaj Agarwal[2]
[1]Assistant Professor, [2]Professor & Head
[1&2] Department of Computer Science & Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India
E-mail: sapyadav08@gmail.com, pankaj7877@gmail.com
(Received 8 January 2020; Revised 20 January 2020; Accepted 4 February 2020; Available online 11 February 2020)

Abstract **- Analyzing online or digital data for detecting epidemics is one of the hot areas of research and now becomes more relevant during the present outbreak of Covid-19. There are several different types of the influenza virus and moreover they keep evolving constantly in the same manner the COVID-19 virus has done. As a result, they pose a greater challenge when it comes to analyzing them, predicting when, where and at what degree of severity it will outbreak during the flu season across the world. There is need for greater surveillance to both seasonal and pandemic influenza to ensure the health and safety of the mankind. The objective of work is to apply machine learning algorithms for building predictive models that can predict where the occurrence, peak and severity of influenza in each season. For this work we have considered a freely available dataset of Ireland which is recorded for the duration of 2005 to 2016. Specifically, we have tested three ML Algorithms namely Linear Regression, Support Vector Regression and Random Forests. We found Random Forests is giving better predictive results. We also conducted experiment through weka tool and tested Zero R, Linear Regression, Lazy Kstar, Random Forest, REP Tree, Multilayer Perceptron models. We again found the Random Forest is performing better in comparison to all other models. We also evaluated other regression models including Ridge Regression, modified Ridge regression, Lasso Regression, K Neighbor Regression and evaluated the mean absolute errors. We found that modified Ridge regression is producing minimum error. The proposed work is inclined towards finding the suitability & appropriate ML algorithm for solving this problem on Flu.**
Keywords**: Epidemics, Influenza Virus, Linear Regression, Support Vector Regression and Random Forests, Zeror, Linear Regression, Lazy Kstar, Random Forest, Reptree**

I. INTRODUCTION

Influenza also known as "the flu", is a viral infectious disease caused by a virus called influenza [1]. Disease like Influenza primarily attacks our respiratory system through viral infection. It can infect your nose, throat, and lungs. Influenza is commonly known as flu. Influenza in most cases gets resolved by its own. However, flu can sometimes cause serious complications particularly with patients having respiratory, heart or diabetes problems and can even be deadly. It causes around 500,000 deaths worldwide each year.

Seasonal influenza spreads more through human contact or when an infected person coughs or sneezes, droplets containing viruses (infectious droplets) are dispersed into the air which can infect others also who encounters droplets. It is advised to cover mouth and nose with a tissue when coughing and wash their hands regularly. Flu occur mainly during changing weather or winters. The time from infection to illness, known as the incubation period, is about 2 days, but ranges from one to four days.

It can be a serious problem for patients with weak immunity and therefore such people are advised to stay home in order to minimize the risk of infecting others in the community, otherwise it can be managed with symptomatic treatment. Vaccination is the most effective way to prevent the disease. Safe and effective vaccines are available and have been in use for many years.

There exist several different types of the influenza virus and moreover they keep evolving constantly in the same manner the COVID-19 virus has done. As a result, they pose a greater challenge when it comes to predicting when, where and at what level of severity the flu will strike during the flu season.

We took the dataset with details of influenza patients of Ireland during 2005 to 2006. The ILI-Denominator feature refers to the actual population size, ILI without the Denominator is the actual cases of influenza, and the numbers denote the age range e.g. 00-04 for population less than 4 years. ILI Number of Cases is the total occurence for the week and is taken as dependent variable for this work. Max Temp, Min Temp and Avg Prec are computed by taking daily and weekly averages of observations. Avg Prec stands for precipitation the data runs from 2005 to 2015.

A.Data Set Contains Following Features

Features= [Country, Year, Week Of Year, Week Start Date, ILI_Denominator00-04,ILI_Denominator05-14, ILI_Denominator15-64,ILI_Denominator65+, ILI_Denominat or Number Of Cases, ILI00-04, ILI05-14, ILI15-64, ILI65+, ILI Number Of Cases, MaxTemp, Min Temp, Avg Prec]

We all know that climatic conditions like temperature, precipitation; humidity plays a major role in the cause of influenza cases. Therefore, we aim to establish the correlation between Influenza Cases in Ireland and the temperature and/or precipitation? We also want to establish which of the correlated feature temperature and/or precipitation plays a bigger role in its cause. We work tried to establish the relationship between the MaxTemp, Min

Temp and Avg Prec. Comparing the graph it is learnt that temperature has a higher correlation with influenza cases than precipitation. Moreover, temperature and week of year are highly correlated. It is also noticed that there are sharp differences between other years and 2009, 2010 and 2011.

II. LITERATURE SURVEY

All went good until "GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%" as Lazer, Kennedy, King, Vespignani reported in the article "The Parable of Google Flu: Traps in Big Data Analysis" on Science(2014). This thing pushed the research community to go further in the prediction analysis of Influenza-Like-Illness from digital data.

Li Zhang *et al*. [3] developed some classification models from a training dataset containing 457 neuraminidase inhibitors and 358 non-inhibitors using random forest and support vector machine algorithms. Neuraminidase has become the foremost target for the treatment of influenza virus. They also used recursive feature elimination (RFE) method to enhance the correctness of the models by selecting the most relevant molecular descriptors. The performances of the models were evaluated by five-fold cross-validation and independent validation. Using the training set, three machine learning models, namely kNN-All, RF-All, and SVM-All, were developed with all molecular descriptors generated by PaDEL-Descriptor. The accuracy of three models namely the kNN-All, RF-All, and SVM-All were measured as 80.7%, 86.3% and 87%, respectively.

Predicting & analyzing epidemics using digital data has become a very hot topic in modern literature. The research paper "Detecting influenza epidemics using search engine query data" by Ginsberg, Mohebbi, Patel, Brammer, Smolinski & Brilliant published by Nature (2009) [4] is considered to be a very important works in research literature.

Aramaki Eiji *et al.* [5] proposed systems that fetch the tweets related to influenza. And then the influenza patients were mined using the SVM classifier model. The data mined from the twitter had 42% negative tweets, in that “influenza” word was included. There is also a very high correlation of 0.89 in the results. Training data sets were having about 5,000 tweets of November 2008. But in excessive news periods, because of bias news, the methods not worked well.

Chen, Po-Fon, *et al*. [6] proposed an algorithm, which is based on SVM classifier and fuzzy measures with high accuracy in immunology prediction. This prediction method had been applied on the HA protein of H2N2 influenza virus for the past three decades to assess the variation of immunogenicity. The strength of immune response of H2N2 viruses HA protein were assessed theoretically. And as a result, decreasing tendency of immune response MH C class II caused increase in probability of H2N2 Virus influenza outbreak.

Broniatowski David A *et al*. [7] proposed the influenza detection algorithm. The predicted influenza rate of the system was correlated with the data of Centres for Disease Control (CDC), and health and mental hygiene of New York City. Supervised classification techniques were used with around 85% accuracy. Proposed algorithm had shown improvements, which is less subtle to the twitter users, as only focuses on the influenza infection reports.

Liu *et al*. [8] applied RNNs LSTM for predicting influenza patterns. They used various novel data sources to forecast influenza patterns, including the geographic spread of influenza, virological surveillance, and the environment, and air pollution, trends in Google. They also discovered several environmental and climatic variables, which are highly correlated with the frequency of ILI.

Zhang *et al.* [9] applied four different LSTM multi-step prediction algorithms to predict influenza outbreaks. They had applied various single-output predictions in a six-layer LSTM framework to attain the greatest precision. The MAPE for the US ILI rates from 2 to 13 steps forward were all less than 15%, with an average of 12.930%.

Yang *et al*. [10] proposed an air quality data analysis and influenza-like illness (ILI) to determine the correlations accurately. They were implemented an integrated platform by a cluster environment, which is based on Hadoop and Spark. The data was collected from 2016 to 2018 in Taichung, Taiwan. The association and visualization between air quality and influenza-like illness was also presented.

Hongxin Xue *et al.* [11] proposed three flu prediction models, to check and verify the cause of spread of flue. The models were based on twitter and US Centers for Disease Control’s (CDC’s) Influenza-Like Illness (ILI) data. They also proposed an amended Particle Swarm Optimization algorithm to optimize the parameters of Support Vector Regression (IPSO-SVR). The IPSO-SVR method was used to predict the regional unweighted percentage ILI (%ILI) events in the US. Following observations were made on the basis of experimental results: 1) flu outbreaks in neighbouring areas also affect the spread of flu in the region; 2) the twitter data supplements with CDC ILI data; 3) the prediction results of IPSO-SVR were better than the prediction results of IAT-BPNN model; 4) the prediction results of IPSO-SVR model 3 were also used to an optimization algorithm that can be used to optimize the SVR parameters as well as for other prediction problem.

Yang, Chao-Tung, *et al.* [12] proposed a regression model and LSTM model to predict rate of influenza-like visits. This study collected data on air pollution and rates of doctor visitation due to ILI. After the air pollution and ILI

treatment rate data were integrated with quality control and time units applied, the data from 2007 to the present were divided into weekly average values for six regions of Taiwan, and the Outbreaks algorithm was used to calculate the influenza-like epidemic. A simple linear regression model was developed to estimate the correlation coefficient between AQI and ILI. According to the model, a significant correlation between PM2.5 and risk of ILI (p value = 0.04) were observed after adjusting for confounders. The deep learning techniques were also used to find correlations between data. Long-term and short-term memory models (LSTM) were used to deal with time-series problems occurring in time-series data. The analysis obtained a valid prediction of the rate of influenza-like visits in the next 4 weeks.

III. PROPOSED WORK

We first preprocessed the data using weka preprocessing module. We first observed the no. of classes grouped by year and found that there are rising trends of cases each year

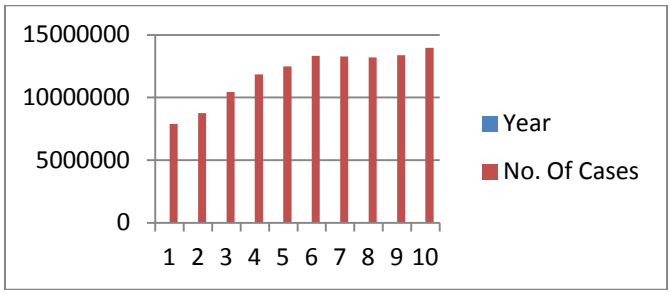

Fig.1 No. of cases vs year

We also plotted the week of year v. Cases, Population Sample, Min Temp

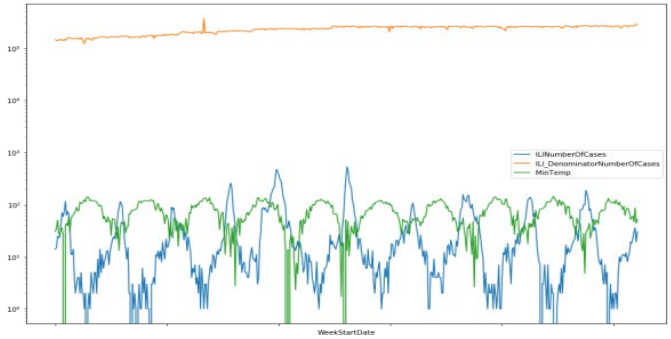

Fig. 2 Week of year v. Cases, population sample, min temp

Following is the plot of Week of Year v. Cases, Precipitation

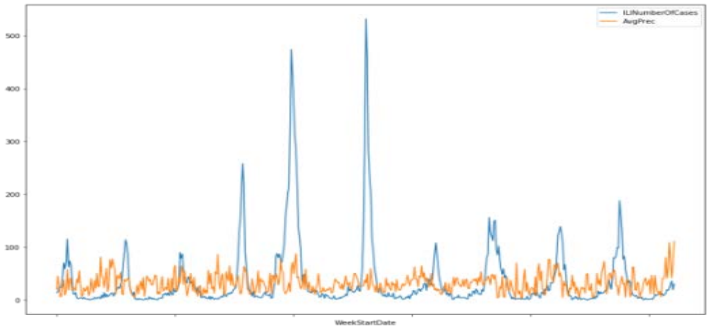

Fig.3 Week of year v. cases, precipitation

Plot of Week of Year v. Cases, Population sample for the different age groups

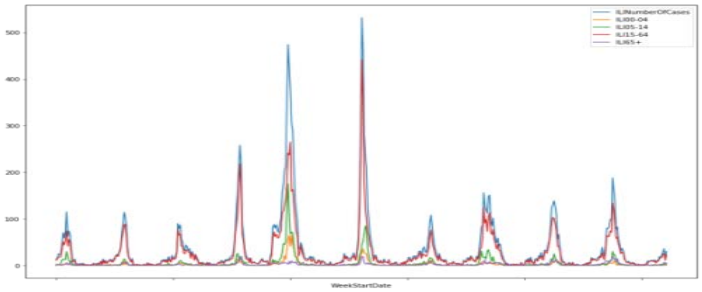

Fig.4. Week of year v. cases, population sample

Following plot between "Week Start Date" VS ["ILI Number of Cases ","Max Temp"," Min Temp"," Avg Prec"] clearly show that the maxTemp has bigger role to play

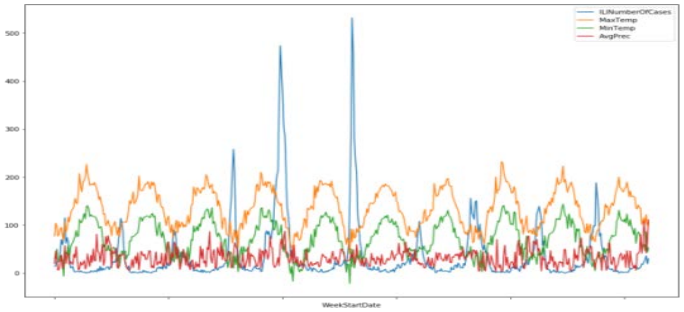

Fig.5 Between "weekstartdate" vs ["ilinumberofcases","maxtemp","mintemp","avgprec"]

All the above correlation plot confirms claim that temperature is highly correlated to influenza cases than precipitation. Heat map showing the co-relations among variables with high correlations in dark pink colors.

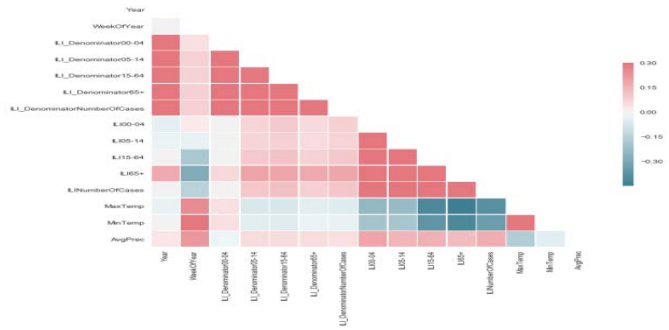

Fig.6 Heatmap

Following is the year-wise plot of the cases. The trend shows a plunge after the 20th week.

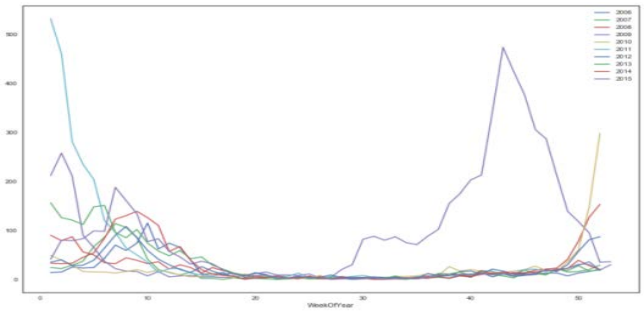

Fig.7 Year-wise plot of the cases

A. Prediction with Linear Regression

We are using 'Week of Year',' MaxTemp',' Avg Prec', 'Min Temp' for this work. We have excluded 'Year' variable from our predictor set though it was giving better accuracies. We are using following two models
Model 1: Week of Year<=1 and Week of Year<=20
Model 1: Week of Year>=20 and Week of Year<=52

Scores for model 1 and 2
Model-1: [-1.56431307 0.50486371 -15.68401103 0.13497511 0.65694866]
Model-2: [-53.95154924 -0.51043076 -0.35925477 0.12535674 -2.70938883]

Root Mean Square value of combined model= 27.586458481105236 In [ ]:

Conclusion - Linear Regression model is not working well on combined model. Two models when combined are not producing good results and therefore we will use one model scheme for rest of the models.

B. Prediction via Support Vector Regression

TABLE I RESULTS OF SVR

K-Fold	Mean Absolute Error	Accuracy of SVR
K-Fold=10	15.306767523002044	-26.3760037981955
K-Fold=20	15.306767523002044	-26.38662511343316
K-Fold=30	15.306767523002044	-26.38981628265173
K-Fold=40	15.306767523002044	-26.316286046463663
K-Fold=50	15.306767523002044	-26.377576668166224

C. Prediction via Random Forests

TABLE II RESULTS OF RANDOM FORESTS

Iteration	Mean Absolute Error	Score
1	28.66538735447559	-3.5347938515976614
2	31.8766848382587	-3.189017032149125
3	29.760134875706296	-3.7871380776945394

D. Other Regression Models

We have also applied other regression models including Ridge Regression, modified Ridge regression, Lasso Regression, K Neighbor Regression and evaluated the mean absolute errors.

TABLE III COMPARISON OF REGRESSION MODELS

Regression Model	Mean absolute error
Ridge Regression	26.22
modified Ridge regression	24.40
Lasso Regression	26.17
K Neighbor Regression (K=3)	28.60
Kernel Regression	28.29

We found that modified Ridge regression model has produced the least Mean absolute error. We had tuned the parameters like alpha, normalize and maximum iteration in the case of modified Ridge regression model.

E. Prediction through Weka Data Mining Tool

Number of Instances=522
Validation Type: 10-fold cross-validation

TABLE IV COMPARATIVE PERFORMANCE OF ALGORITHMS USING WEKA

Algorithm	Mean absolute error	Root mean squared error	Relative absolute error (%)
ZeroR	37.9088	64.5484	100
Linear Regression	31.4546	59.1913	82.974
Lazy IBK	34.6341	81.56	91.361
Lazy KStar	30.1759	63.6694	79.6014
Random Forest	31.0259	31.0259	81.8436
REP Tree	31.0654	63.8066	81.9476
Multilayer Perceptron	37.8812	64.1079	99.9273

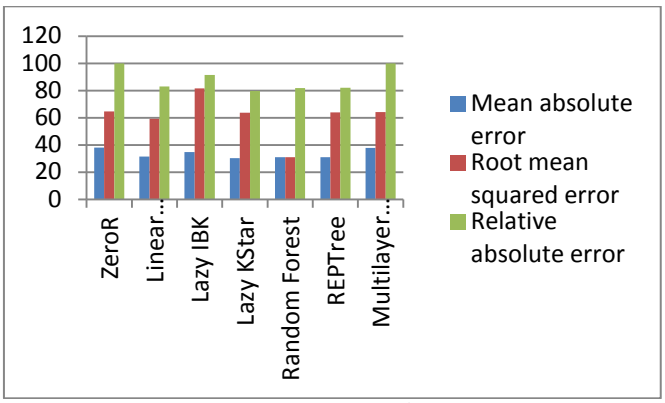

Fig.8 Graphical comparison of algorithms

IV. CONCLUSION

The proposed work is inclined towards finding the suitability & appropriate algorithm for solving this problem on Flu. Random forest shows promising results in comparison with SVR, Linear Regression, ZeroR, IBK, Multilayer Perceprton, REP Tree, and KStar. Among the regression models modified Ridge regression seems to produce minimum error.

REFERENCES

[1] "Influenza (Seasonal)". *World Health Organization (WHO).* 6 November 2018. Archived from the original on 30 November 2019. Retrieved 30 November 2019.

[2] "Key Facts about Influenza (Flu)". Centers for Disease Control and Prevention (CDC). 9 September 2014. Archived from the original on 2 December 2014. Retrieved 26 November 2014.

[3] Li Zhang, Haixin Ai, Qi Zhao, Junfeng Zhu, Wen Chen, Xuewei Wu, Liangchao Huang, Zimo Yin, Jian Zhao, and Hongsheng Liu, "*Computational Prediction of Influenza Neuraminidase Inhibitors Using Machine Learning Algorithms and Recursive Feature Elimination Method," Virology*, Vol.352, pp. 418-426, 2006.

[4] Ginsberg, Mohebbi, Patel, Brammer, Smolinski & Brilliant, "Detecting influenza epidemics using search engine query data" published by Nature, 2009.

[5] Aramaki Eiji, Sachiko Maskawa, Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

[6] Chen, Po-Fon, *et al*. "Theoretical assessment of immunogenicity variation on the HA protein of H2N2 influenza virus by using fuzzy integral and SVM classifier." *International Conference on Machine Learning and Cybernetics*. Vol. 2. IEEE, 2011.

[7] Broniatowski David A, Michael J Paul, Mark Dredze. "National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemics", *PloS one*.Vol.8, No.12,2013; e83672

[8] Liu L, Han M, Zhou Y, Wang Y., "LSTM recurrent neural networks for influenza trends prediction", *In: International symposium on bioinformatics research and applications*. Springer, Cham, pp. 259–264, 2018.

[9] Zhang J, Nawata K, Multi-step prediction for influenza outbreak by an adjusted long-short term memory. *Epidemiol Infect* Vol.146, No.7, pp.809–816, 2018.

[10] Yang CT, Chen CJ, Tsan YT, Liu PY, Chan YW, Chan WC, "An implementation of real time air quality and influenza-like illness data storage and processing platform", *Comput Hum Behav*,2018

[11] Xue, H., Bai, Y., Hu, H., & Liang, H., "Regional level influenza study based on Twitter and machine learning method." *PloS one* 14.4, 2019.

[12] Yang, Chao-Tung, *et al*. "Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources." *The Journal of Supercomputing*, 2020.