Predict H1N1 and Seasonal Flu Vaccines

Jawad Adil
dept. of Computer Science,
Griffith College
jawad.adil@student.griffith.ie

Maureen Chianumba
dept. of Computer Science,
Griffith College
maureenchinedum.chianumba@student.griffith.ie

Guilherme Coelho
dept. of Computer Science,
Griffith College
Guilherme.coelhodeazevedo@student.griffith.ie

***ABSTRACT* - In the spring of 2009, a novel influenza A (H1N1) virus emerged and was first detected in the United States which quickly spread around the world. This virus caused worldwide deaths between 150,000 to 600,000 in a calendar year. It was so severe that the World Health Organisation (W.H.O) declared it a global pandemic. A vaccination program against H1N1 influenza was introduced in many countries but there was low uptake in both the general population and health professionals, so a National 2009 H1N1 Flu Survey (NHFS) was conducted in the US to check the status of vaccination. The data from the flu survey is used in this paper to predict the possibility of a person receiving H1N1 and Seasonal Flu Vaccine.**

Keywords - H1N1, Seasonal, Vaccine, influenza, Virus, Classification.

I. INTRODUCTION

Influenza was the first infectious respiratory disease to be monitored globally. It is a viral disease that crosses regions quickly through interpersonal communications and it is very fatal among certain populations. Seasonal influenza viruses and H1N1 are proteins found on the surface of the virus. The human immune system protects the body against influenza infection by producing antibodies that can recognize these proteins. However, the influenza virus mutates frequently, including at sites that affect the immune system's ability to detect the virus.[1]

H1N1 according to virology is a subtype of Influenza. It is a virus that is also written as (A/H1N1). Influenza 'A' virus was also a common reason for Influenza (flu) in 2009 - 2010 and is also associated with the great plague of the 20th century that devastated Spain (1918-1920). The H1N1 virus was declared a pandemic in June of 2009 by W.H.O. which was the cause of approximately 150,000 to 600,000 deaths worldwide recorded in the tenure of 20 months. Seasonal flu or Flu season, on the other hand, is quite common and not that deadly. It occurs in recurring periods. It also mostly occurs in cold weather. The three main virus families that can be pinpointed for the seasonal flu are Influenzavirus A, B, and C [2]

Seasonal influenza affects the worldwide population every year, while pandemic influenza is an unpredictable threat.[3]. Following the WHO recommendations for the prevention and control of influenza, more than 40% of global nations currently include vaccinations in their public health policies [4] and [5]. However, a large contest has developed about the evidence used, transparency of the decision-making process, and the role of the institutions involved [6], [7], and [8].

The prevention of influenza is seriously dependent on the appropriate monitoring of flu outbreaks. Advanced warnings can help to manage the spread of the disease, thereby reducing the number of cases and fatalities caused by an epidemic. Even though there was a rise in the number of influenza studies after the last pandemic (H1N1), investigations regarding influenza vaccination policymaking are still scarce.

In this paper, six machine learning models were implemented to predict the possibility of a person taking the H1N1 and Seasonal vaccine. The models used include Random Forest, Logistic Regression, SVM, Decision Trees (J48/C4.5), Stochastic gradient descent (SGD), and KNN.

This paper is divided into the following section: Sect. I give an overview of the subject and the abstraction of this paper. Sect. II describes seasonal influenza and H1N1, demonstrating situations in the past when society didn't have enough technology to create vaccines, creating threats worldwide. Sec. III contains papers where predictions and forecasts about seasonal influenza and H1N1 are suggested. In addition, there are papers related to vaccination in both viruses. Sec. IV describes the methodology used in this paper. It introduces the dataset used in the research, preprocessing done, generation of training and test datasets, and hyperparameters tuning. Sec. V presents the evaluation used in the models. Sec VI presents the algorithms used to create models and the results obtained from each model. Sec VII contains the conclusion and future work. Sec VIII contains the references used in this paper.

II. RELATED WORK

Jhaveri J. [9] In his paper titled "Flu Shot Learning: Predict H1N1 And Seasonal Flu Vaccines" described the different machine learning algorithms and artificial neural network models used to predict the probability of taking the vaccine. The result showed that Artificial Neural Network was the best method with 2 hidden layers having an accuracy of 82% and 86%. Random forest and SVM also yielded good results except for the logistic regression which was the least performing model with an accuracy of less than 70% in both H1N1 flu and seasonal flu vaccination.

A paper was done by Werth R. [10]. titled "Using classification models to predict vaccinations", he used

various deep learning models for predicting the probability of an individual having taken the H1N1 vaccine or seasonal flu vaccine. He focused on the false positive because of how imbalanced the dataset was so he couldn't rely on the accuracy alone to determine how good the model was. He compared the results of various machine learning models to improve the precision of prediction with logistic regression, random forest, decision trees along with the gradient boosting algorithms and got a recall of 95% in both logistic and random forest.

There are other similar research paper topics that have been published. Bish et al. [11] in their paper titled "Factors associated with uptake of vaccination against pandemic influenza" described the psychological and demographic factors influencing the spread of the H1N1 vaccine and also did some work based on deep learning models. The results showed that people who thought others wanted them to be vaccinated were more likely to do so and people getting their information about vaccination from official health sources.

Xue et al. [12] worked on multiple regression models and artificial neural networks to monitor flu activity. In their paper titled "Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network" focuses on a wide variety of models that rely on data and values from the Google Flu Trends (GFT) and the Centres for Disease Control (CDC) to predict the characteristics of the flu. The results showed that the GFT+CDC regression model is preferable for monitoring influenza activity compared to the GFT and CDC regression models.

Venna et al. [13] in their paper titled "A Novel Data-Driven Model for Real-Time Influenza Forecasting", is used to predict flu in real-time. The article proposes a new data-driven machine learning method using long-term multi-stage memory-based forecasting to predict flu. The results showed that the LSTM-based deep learning method showed better results than time series forecasting methods.

Nieto-Chaupis et al [14] In their paper titled "Face To Face with Next Flu Pandemic with a Wiener- Series-Based Machine Learning: Fast Decisions to Tackle Rapid Spread" explained the Wiener model used to increase optimization, efficiency, and performance to find the spread of seasonal flu. The result showed that the Bessel function might be correlated to the number of infected cases while kernels are the performance of the applied task.

Joshi et al [15] in their paper titled "Shot or Not: Comparison of NLP Approaches for vaccination behavior detection" focuses on the rule-based, statistical and deep learning approach to detect vaccination behavior. The result showed that an ensemble of statistical classifiers using pre-trained language models and LSTM classifiers obtained better performance for vaccination behavior detection.

Yang et al [16] in their paper titled "Influenza-like illness prediction using a long short term memory deep learning model with multiple open data sources" shows influenza-like illness and predicts the outbreak of influenza in a region. The result showed that the LSTM model generated a very good performance and the TensorFlow library and Keras tools were used to implement neural network models for the experiment.

Mabrouk et al [17] in their paper titled "A chaotic study on pandemic and classical (H1N1) using EIIP sequence indicators" used the K-means clustering algorithm to classify the two different types of influenza along with the correlation dimension. The result showed that there is no significant difference between the two types of H1N1 based on the p-values of the t-test performed and the non-linear features performed well with high accuracy.

III. METHODOLOGY

A. Dataset

In search of a useful dataset that has good attributes for classification problems, the H1N1 and Seasonal flu vaccine data provided by DrivenData which comes from the national 2009 H1N1 flu survey (NHFS) was selected. The main goal was to predict whether a person will get an H1N1 or seasonal flu vaccination or not. The data was in raw form that needed to be processed. The dataset had 36 attributes and 2 marked labels in separate files with a total of more than 26 thousand instances. Out of those 36 attributes, 24 were numeric, 8 were categorical and 4 simple string attributes.

B. Pre-Processing and EDA

The first step taken was to visualize and analyze the data. The main problem was divided into two parts. First, predict the H1N1 flu vaccine and second was to predict seasonal flu. Moreover, out of 26 thousand instances and 36 columns, there were some attributes with missing values. The columns with and without incorrect values were noted. The data and labels were in separate files which were to be combined into a single file to be used in weka. The following paragraph will elaborate on how we dealt with these issues.

The second step was to pre-process the data which was organized in different sub-steps. The first sub-step was to combine the labels and data into one file using Microsoft Excel and the next sub-step was to remove 2 unnecessary columns "Employment occupation" and "Employee Industry" that had more than half of the missing values from a total of 26 thousand instances using Weka. The next thing was separating the data for H1N1 and Seasonal vaccines by removing seasonal vaccine-related columns first and saving the arff file for H1N1 and doing the same thing by removing H1N1-related attributes to get a dataset for seasonal vaccines. By doing this, there were two datasets for both sub-problems which are already defined in the previous paragraph.

The next Step was to deal with the missing values. Several columns had around 4000 missing values and they were going to affect the results if replaced using some filters like mean median or mode. So those instances with a large number of missing values were removed. For the columns with less than 200 missing values, the "ReplaceMissingValues" filter of weka was used. Doing this resulted in around 21 thousand instances without missing values. The second technique was to remove those instances

that had 1 or more missing values. To remove those instances, the "RemoveMissingValues" filter for numeric values was used. For the categorical or string variables, weka doesn't have an option to directly deal with those values. A trick was used here and the data was sorted and a "RemoveRange" filter was applied to remove the missing values which are placed at the end of the column due to sorting. This step was repeated for every column in the dataset that had a string or categorical values. This process was done for both datasets i.e. for H1N1 and seasonal flu but the attributes were different for both datasets based upon the feature importance.

For feature selection, Naive Bayes and Bagging were used within weka. The features with 0 merit/rank were removed. This resulted in a total of 29 attributes for H1N1 and 25 attributes for seasonal vaccine predictions.

In the evaluation of the dataset, it's possible to notice the difference between shots taken in H1N1 and seasonal vaccines. There are a bigger number of seasonal vaccinations compared to H1N1. This indicates that possibly there are some key attributes to predict the H1N1 vaccine.

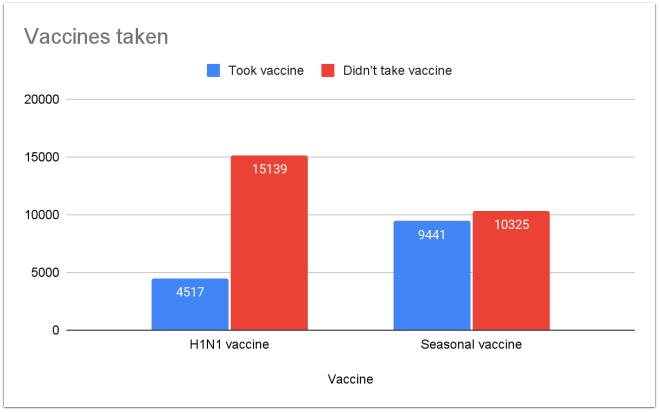

Figure 1: Ratio of vaccines

When analysing education, there are 4 categories: less than 12 years studying, 12 years studying, college graduate and some college. Most surveyed were college graduates:

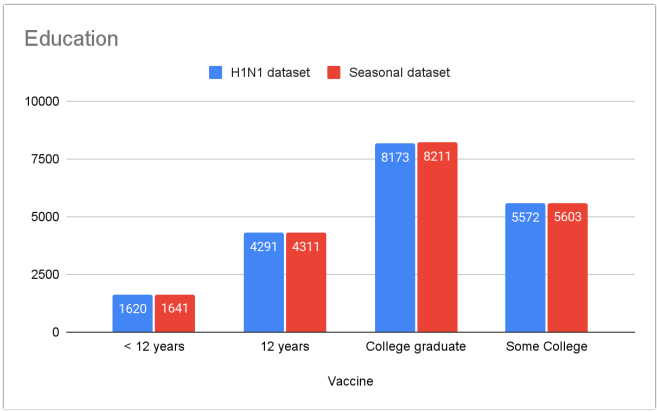

Figure 2: Education levels

When analysing sex, there were more women in the survey than men:

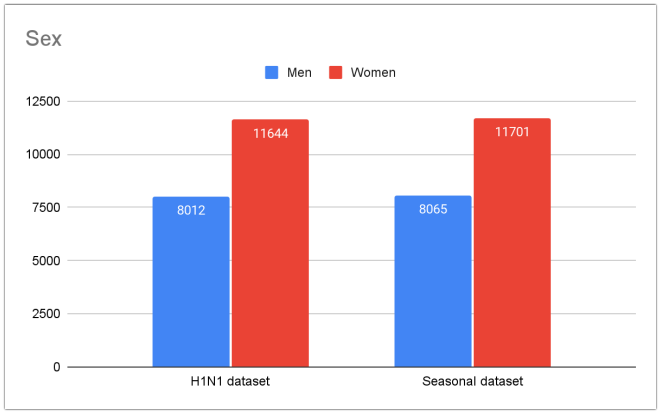

Figure 3: Distribution of sex attribute

In the age group attribute, the surveyeds were grouped in 5 categories: 18 - 34 years, 35 - 44 years, 45 - 54 years, 55 - 64 years and 65 years or older. The age group 65+ years is the group containing the biggest number of people, followed by the other 2 older groups. These groups tend to worry more in getting vaccinated due to the vulnerability of their immune system.

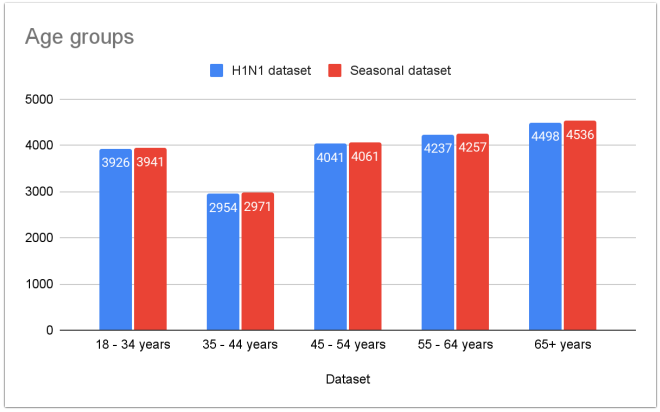

Figure 4: Age group comparison

C. *Getting Training and Testing files*

After the values are cleaned, the dataset was divided into train and test datasets. The "RemovePercentage" filter was used with 25% to get 75% for training. Then the same filter was applied with the "invertSelection" attribute to get 25% for test data. The same process was repeated for the seasonal dataset. At the end of this process, 8 new files were created. The first four with removed missing values containing 2 training and 2 testing files for both H1N1 and seasonal flu. The other four with a combination of removed and replaced instances for missing values. The division of these 4 files is the same as in the 1st case i.e. 2 training and 2 testing sets for both H1N1 and seasonal flu. The whole pre-processing and the splitting of the dataset is summarised in the image below.

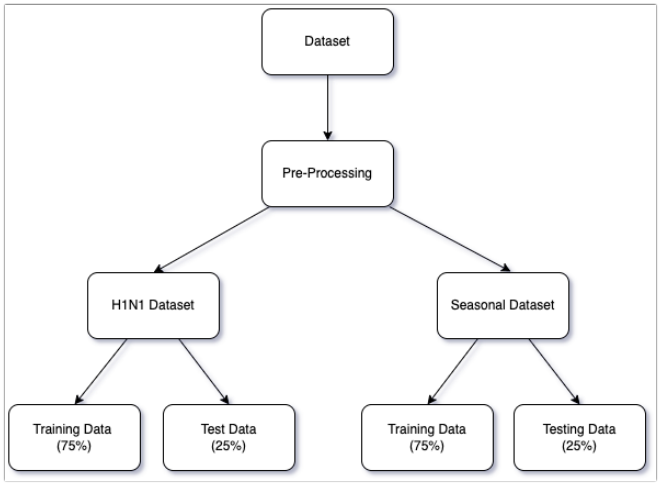

Figure 5: Illustration of getting training and testing files

D. Hyperparameters Tuning

Models like decision trees, logistic regression, support vector machines and others used in this analysis have some parameters that need to be adjusted before we actually run the model. This adjustment is done to get more accuracy. Since we are using multiple models to test which one performs better, We have adjusted parameters for every algorithm we used. Then results were compared and the parameters which gave best accuracy were chosen for every algorithm. The selected parameters are given in table 6.1.

IV. Evaluation of the Results

There are different evaluation matrices used in weka by default. Some important are Area under the curve (AUC), F-Measure, Matthews Correlation Coefficient (MCC), Precision/Recall, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Also, we get the output of the confusion matrix with classes to be predicted which is used to get the values of most of the evaluation matrices named above.

In our case, we used the default evaluation matrices of weka. Talking about AUC ROC (Area under the curve), higher the value of AUC, the better the model is predicting the values. It lies between 0 and 1, 0 being the worst and 1 being the best. Precision is the percentage of relevant results and Recall is related to the results correctly classified by an algorithm.

The confusion matrix is used to calculate the Accuracy (all correct / all), Precision (true positives / predicted positives) and Recall (true positives / all actual positives). Given as:

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

Once we have precision (P) and Recall (R), we can calculate F-Measure using this formula:

$$F - Measure = \frac{2 * P * R}{P + R}$$

F - Measure is the combined measure of precision and recall obtained by giving equal importance to both of the values. The range for F - Measure is 0 to 1 and 0 being the worst case and 1 the best case.

V. Results

A. Decision Tree

The first model used was decision tree (J48/C4.5). Using the default parameters of weka, i.e. coincidence factor = 0.25 and minNumObjects = 2, it got an accuracy of 83.6182% for H1N1 and 76.1028%. After tuning the hyperparameters, there was a slight increase and the new accuracy was 84.1677% for H1N1 and 77.3978%. Comparison of the accuracies before and after tuning hyperparameters is given below.

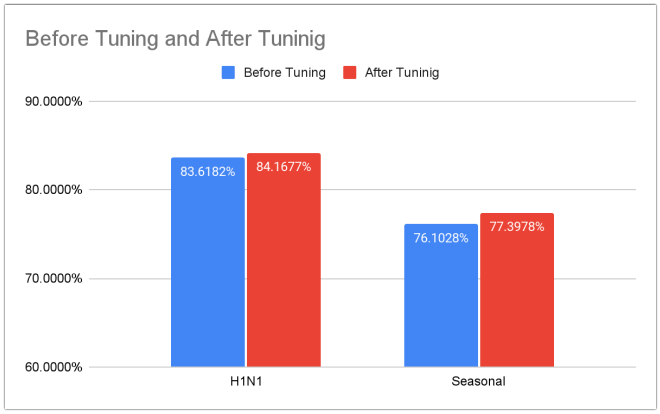

Figure 6: Accuracy comparison of decision tree before and after tuning

B. Random Forest

Random Forest was the second model used for this research. It achieved an accuracy of 84.2898% for H1N1 and 77.2966% for seasonal vaccines. After hyperparameters tuning i.e., depth = 70 and Iterations = 500, it achieved a small improvement in the accuracy. The new accuracy was 84.4729% for the H1N1 vaccine and 77.6002% for the seasonal vaccine. Following chart demonstrates the difference between the accuracies.

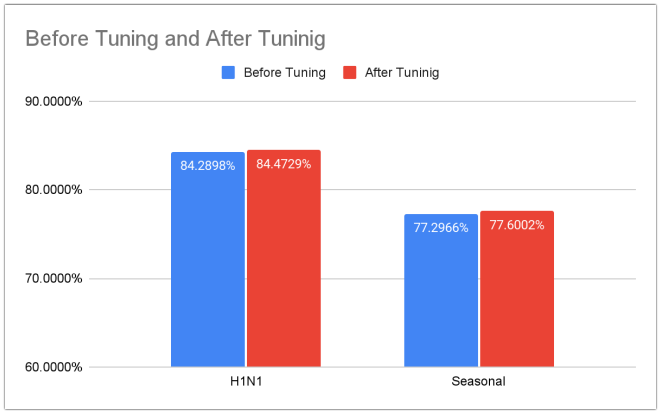

Figure 7: Accuracy comparison of random forest before and after tuning

C. *Logistic Regression*

The third model used is the logistic regression which is the best performing algorithm. The accuracy of Logistic Regression before tuning its parameters for H1N1 was 85.0224% and 78.6524% for seasonal. It achieved an accuracy gain by tuning the parameters. The best parameters found for this algorithm includes weight trimming for logit boost to 0.1 with (AIC) set to true By doing this, AIC checks for the number of iterations to stop for logit boost. The accuracy after tuning reached upto 85.1852% for H1N1 and 78.9964% for seasonal. A comparison before and after tuning is given below.

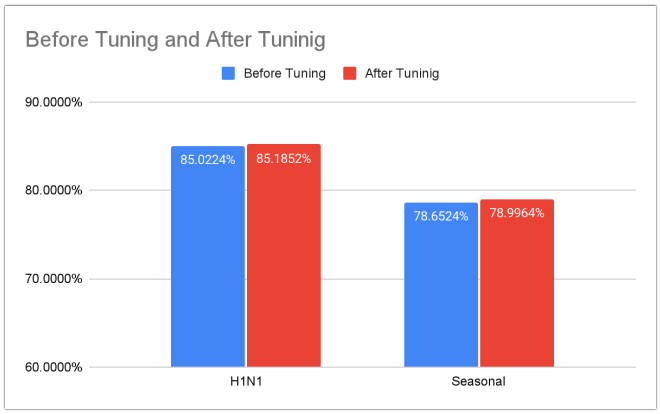

Figure 8: Accuracy comparison of logistic regression before and after tuning

D. *Support Vector Machine (SVM)*

The fourth model tested was Support Vector Machine (SVM). This is the second-best performing model in our research even without changing the kernel. The accuracy with PolyKernal was 84.9206% for H1N1 and 78.4500% for Seasonal Vaccines. The highest accuracy was achieved by the RBF kernel. The new accuracy was 84.9817% for H1N1 and 78.8142% for seasonal vaccines. Following chart illustrates the difference between 2 kernels used.

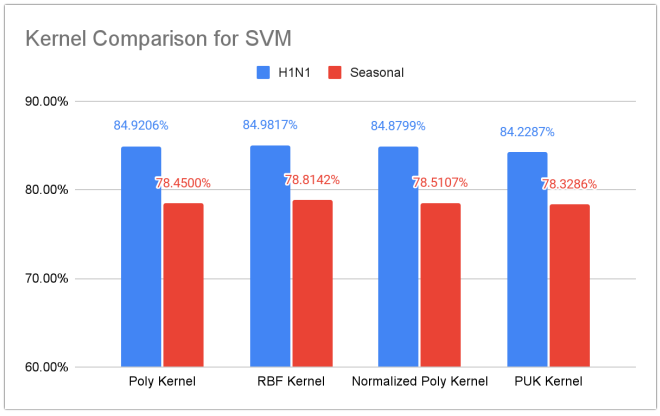

Figure 9: Accuracy comparison of support vector machine before and after tuning

E. *SGD*

The fifth model used was SGD. It gave an accuracy of 84.7375% for H1N1 and 77.8227% for seasonal vaccines on default weka parameters. The hyperparameters were adjusted to epochs=500 and loss function=Log Loss to get the best out of this algorithm. The highest accuracy gained was using the above parameters. 84.9206% of H1N1 and 77.9846% instances were correctly identified after this tuning. Following bar chart shows the difference.

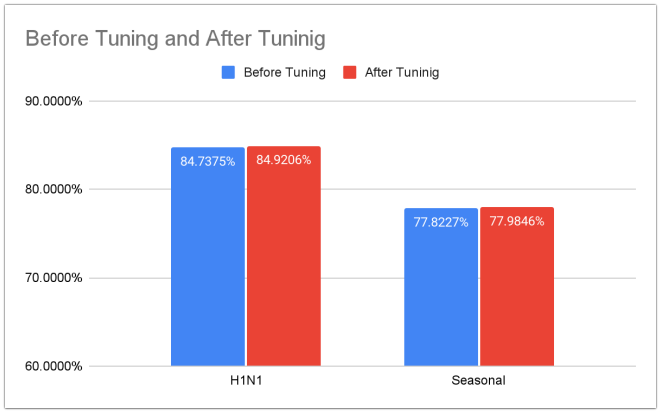

Figure 10: Accuracy comparison of SGD before and after tuning

F. *KNN*

The last algorithm tested was KNN. Initially, this model didn't perform well and gave 78.022% accuracy for H1N1 and 71.3072% for seasonal vaccines. The parameters were adjusted to get more accuracy. The impact of adjusting parameters resulted in the gain of accuracy for both seasonal and H1N1 vaccines. The adjusted parameters for this algorithm were Neighbours = 10, Algorithm = LinearNNSearch, Distance Function = Manhattan Distance. The new accuracy was 82.7228% and 76.4063% for H1N1 and seasonal vaccines respectively. Following chart is added to illustrate the impact of adjustment of parameters.

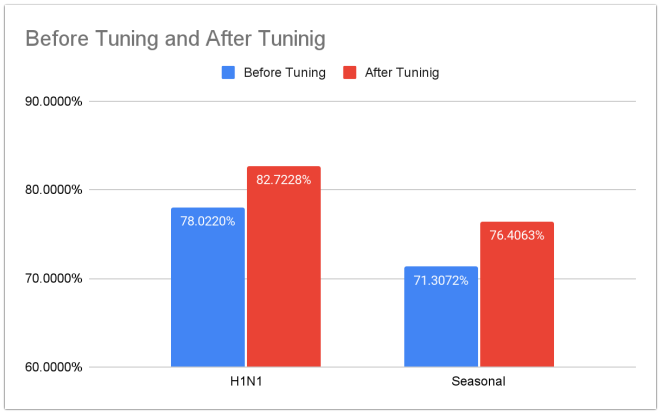

Figure 11: Accuracy Comparison of KNN before and after tuning

The second highest accuracy was gained by Support Vector Machine (SVM/SMO) for both seasonal and H1N1. So, we tried stacked and voted multiple classifiers using both Logistic Regression and SVM keeping Logistic Regression as meta classifier. The results were not good, so we skipped the stacking of these algorithms.

Following chart shows the comparison of the highest accuracies gained from every model used in this paper.

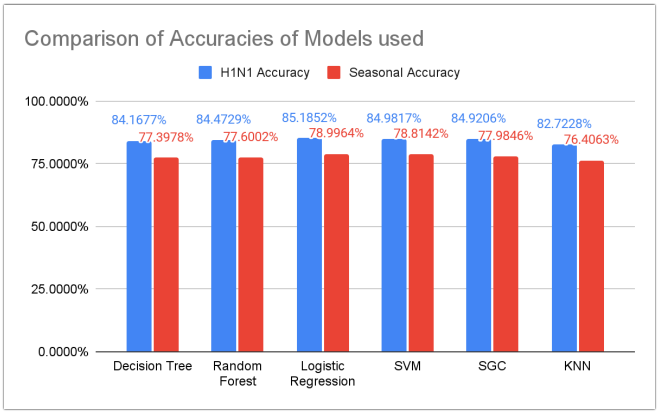

Figure 12: Accuracy comparison of all models

In this chart above, the highest accuracy for both H1N1 and seasonal vaccine prediction was achieved by the Logistic Regression. Although SVM also gave good results. Stacking both these algorithms slightly increased the accuracy from what we got from the SVM, but the resulting accuracy was still less than the accuracy of Logistic Regression. Using default parameters of weka for logistic regression, the accuracy for H1N1 was 85.0631% and 78.5512% for seasonal. An increase in accuracy was achieved by fine tuning the hyperparameters and this percentage was raised up to 85.1852% for H1N1 and 78.9964% using the same Linear Regression model. Hence, it can be used in such cases to draw good results.

VI. Conclusions and Future Work

Influenza is a deadly disease that causes a massive economic loss and severe effect to societies across the globe yearly and being vaccinated can help to further reduce the spread of this virus.

The goal is to predict the possibility of a person getting the H1N1 and seasonal flu vaccine based on a series of questions asked to that person. We tried 6 different algorithms and got some good results. Linear Regression predicted the highest number of the instances correctly compared to other models. So, to conclude this paper, we can say that the linear regression model is proven to be best for prediction in our case.

References

1. Perofsky A.C and Nelson M.I. (2010). Seasonal Influenza: The Challenges of vaccine strain selection (Online) https://elifesciences.org/articles/62955 [Accessed 27 November 2021].

2. Charles Patrick Davis. “Flu (influenza, conventional, H1N1, H3N2, and bird flu [H5N1]) facts” (Online) https://www.medicinenet.com/influenza/article.htm [Accessed 30 November 2021].

3. WHO World Health Organisation: fact sheet seasonal influenza (Fact sheet No. 211), WHO Influenza Seas WHO (2009).

4. A. Palache Seasonal influenza vaccine provision in 157 countries (2004–2009) and the potential influence of national public health policies Vaccine, 29 (2011), pp. 9459–9466

5. E. Miller (2011). Report from the SAGE working group on influenza vaccines and immunizations.

6. W.E.P. Beyer, J. McElhaney, D.J. Smith, A.S. Monto, J.S. Nguyen-Van-Tam, A.D.M.E.Osterhaus (2013). Cochrane re-arranged: support for policies to vaccinate elderly people against influenza Vaccine pp. 6030-6033.

7. T. Jefferson, C. Di Pietrantonj, L.A. Al-Ansary, E. Ferroni, S. Thorning, R.E. Thomas. (2010). Vaccines for preventing influenza in the elderly (review) Cochrane Database Syst. Rev. p. CD004876.

8. J.A. Knottnerus Influenza vaccination in the elderly: current evidence and uncertainties Journal of Clinical Epidemiology, 62 (2009), pp. 675–676.

9. Jhaveri J. (2020) Flu Shot Learning: Predict H1N1 And Seasonal Flu Vaccines (Online) https://vesitaigyan.ves.ac.in/flu-shot-learning-predict-h1n1-and-seasonal-flu-vaccines/ [Accessed 30 November 2021]

10. Werth (2020). Using classification models to predict vaccinations (Online) https://medium.com/analytics-vidhya/using-classification-models-to-predict-vaccinations-f71d1c43bec7 [Accessed 30 November 2021]

11. A. Bish, L. Yardley, A. Nicoll, and S. Michie (2011)“Factors associated with uptake of vaccination against pandemic influenza: A systematic review,Vaccine,doi:10.1016/j.vaccine.2011.06.107.

12. Xue, Hongxin & Bai, Yanping & Hu, Hongping & Ldfs, Hdsajkkd. (2017). Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network.IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2771798.

13. S. R. Venna, A. Tavanaei, R. N. Gottumukkala, V. V. Raghavan, A. S. Maida, and S.Nichols (2019)“A Novel Data-Driven Model for Real-Time Influenza Forecasting,” IEEE Access, doi: 10.1109/ACCESS.2018.2888585.

14. Nieto-Chaupis, Huber.(2019). Face To Face with Next Flu Pandemic with a Wiener-Series-Based Machine Learning: Fast Decisions to Tackle Rapid Spread. 0654-0658 (Online) https://ieeexplore.ieee.org/document/8666474 [Accessed 4 December 2021]

15. Joshi A, Dai X, Karimi S, Sparks R, Paris C and Macintyre C.R (2018). Shot or Not: Comparison of NLP Approaches for vaccination behaviour detection. (Online) https://www.researchgate.net/publication/334115834_Shot_Or_Not_Comparison_of_NLP_Approaches_for_Vaccination_Behaviour_Detection [Accessed 12 December 2021]

16. Yang C, Chen Y, Chan Y, Lee C, Tsan Y, Chan W, and Liu P. (2020). Influenza-like illness prediction using a long short term memory deep learning model with multiple open data sources (Online) https://link.springer.com/article/10.1007/s11227-020-03182-5 [Accessed 13 December 2021]

17. Mabrouk M, Marzouk S. (2010). A chaotic study on pandemic and classical (H1N1) using EIIP sequence indicators (Online) https://ieeexplore.ieee.org/document/5645882/authors#authors [Accessed 13 December 2021]