Using NHFS Data to Predict Influenza and Seasonal Vaccination Rates: A Comparison of Different Machine Learning Algorithms

Jiya Benny
School of Computer Science,
The University of Nottingham, NG8 1BB
Nottingham, United Kingdom
psxjb8@nottingham.ac.uk

Ann Mary Thomas
School of Computer Science,
The University of Nottingham, NG8 1BB
Nottingham, United Kingdom
psxat10@nottingham.ac.uk

***Abstract*—The 2009 Swine Flu outbreak spread the epidemic, and great efforts were made to monitor, avoid, and immunize to stop the spread. In an attempt to analyze and predict the two types of vaccine namely H1N1 Vaccine and Seasonal Vaccine,this study seeks to find an appropriate method to predict the probability of these two vaccines based on the different responses collected from the people considering the various factors related to the Flu. After performing an initial analysis to study the relationship between various features obtained from the survey using various plots and correlation analysis. It then compares the performance of different machine learning models, namely logistic regression, Random Forest classifier, Support vector machines, random forests, Gradient Boosting classifier, Naïve Bayes Classifier, K Nearest Neighbor which are different machine learning approaches to predict the probability of how likely individuals are to receive their H1N1 and seasonal flu vaccines. We perform cross validation and hyper tuning to improve the accuracy of some models. In the end, we observe the support vector machine model performs best in predicting the probability of H1N1 vaccine and gradient boosting for seasonal vaccine. We also gain certain additional insights about the relationship between various features for predicting the probability.**

***Index Terms*—H1N1, Prediction, Classification, Machine Learning, Support Sector Machines, Random Forest, Decision Tree, Logistic Regression, Gradient Boosting Classifier, Cross Validation and Hyper Tuning.**

I. Introduction

In the spring of 2009, there was an outbreak of severe pneumonia associated with a novel swine origin influenza A virus, commonly called as swine influenza. The outbreak was first reported in Mexico and quickly spread to the rest of the globe, causing a global pandemic. The H1N1 virus is a subtype of influenza A virus and is considering to be derivational from a combination of avian, swine and human influenza viruses. The World Health Organization (WHO) declared this as a global pandemic in June 2009. The virus has affected millions of people around the universe and initiated great efforts to control its spread and prevention.

The main reason for this global pandemic were the H1N1 virus was a new strain that was not previously identified in human being. The virus may have under gone genetic mutation, creating new strain that could easily spread from human to human. Another reason was due to the advancement in transportation international travel and trade facilitate the the spread of this disease. The H1N1 virus entered United States through infected person travelled from affected regions. Another important reason might be due to the lack of immunity, Since the H1N1 was a new strain people had only little pre-existing immunity towards it. This made a large portion susceptible to infection.

By October 5, 2009, a separate monovalent vaccine was developed and made available to the public to combat the H1N1 strain, as it emerged too late to be included in the trivalent seasonal influenza vaccine for the 2009-10 season. ssThe NHFS collected data on the uptake of both the H1N1 and seasonal influenza vaccines from a target population comprising all persons in the United States aged 6 months and older.

Recently, various machine learning, deep learning and Artificial Neural Network techniques are used to analyze more about this pandemic and for suggesting future prevention method. This paper is a comparative study of some machine learning techniques to predict the probability of vaccine from the selected attributes. Six supervised and unsupervised machine learning approaches have been developed to predict how likely individuals are to receive their H1N1 and seasonal flu vaccines with maximum accuracy.

A. Dataset

The data sets used in this research is taken from the open-source driven data repository, which is available as a comma-separated (CSV) downloadable file having separate file for training and testing. We performed the analysis using Python Jupyter notebook.Generally, there are two types of vaccine:

1) H1N1 vaccine- Vaccines stimulate the immune system to produce antibodies that can recognize and neutralize the H1N1 virus, reducing the risk of infection and its associated complications.
2) Seasonal Vaccine- Designed to provide immunity against the most common influenza strains and reduce the chance of infection and serious illness.

In this dataset, we’ll be considering H1N1 and seasonal vaccines probability. These two are the target variables. Both are binary variables:0=No;1=Yes. Some respondents didn't get either vaccine, others got only one, and some got both. The dataset has 36 features giving a detailed information about the factors considered in the survey like h1n1_concern (level of concern about the H1N1 flu), behavioral_wash_hands (frequently washed hands or used hand sanitizer), doctor_recc_h1n1 (H1N1 flu vaccine was recommended by doctor), opinion_h1n1_vacc_effective (respondent's opinion about H1N1 vaccine effectiveness), and many more. The data set contains majority of features as binary variables and some categorical variables.

B. Aim

The purpose of this study is to develop effective machine learning approaches for predicting the likelihood how likely individuals are to receive their H1N1 and seasonal flu vaccines using multiple machine learning algorithms such as Support Vector Machines, Random Forest, Logistic Regression, Gradient Boosting and Decision Tree classifier. The performance of each classifier will be evaluated in terms of accuracy, training process and testing process. Hyper tuning and cross validation are performed to get more accurate result.

I. Literature Review

There is many research paper related to this topic. Nieto-Chaupis, Huber. (2019) [1]. In this paper “Face to Face with Next Flu pandemic with Wiener-Series-Based Machine Learning: Fast Decisions to Tackle Rapid Spread” describe a Machine learning model to increase the performance, efficiency and optimization to find the spread of seasonal flu and Chinh et al. (2010)[2] in this paper “A possible mutation that enables the H1N1 influenza A virus to escape antibody recognition” studies about the method like Phylogenetic Analysis of Pandemic Strains, Molecular docking for the predicted epitopes and they concluded through their studies B cell epitope helps the virus from antibody recognition and gives vaccine against H1N1. Bao et al. [3] find in their paper "Influenza-A Circulation in Vietnam through Data Analysis of Hemagglutinin Entries" provided with influenza virus data sources then used for the analysis of influenza virus and Hu et al. [4], in their publication Computational Study of Interdependence Between Hemagglutinin and Neuraminidase of Pandemic 2009 H1N1" explained Information model Spectrum and sequence data to study between Hemagglutinin and neuraminidase of H1N1. Stalder et al. [5] in the famous paper “Tracking the flu pandemic by monitoring the social web” which is related to collecting data from Twitter and other official health reports provides timely and good information about the epidemic. Wiriyachaiporn et al. [6], In their research paper "Rapid influenza an antigen detection using carbon nano string as the label for lateral flow immunochromatographic assay” showed a preparation of allantoic fluid infected with influenza A A virus conjunction of CNS to antibody and about the evaluation of CBNS-MAB using LFIA are used in the detection of nanoparticles Ma et al.[7] published a paper on the topic “An integrated passive microfluidic device for rapid detection of influenza a (H1N1) virus by reverse transcription loop-mediated isothermal amplification (RT-LAMP)” depicted the loading about magnetic bread and loading of virus, virus capture, collection of magnetic breads, virus particle lysis, removal of excess waste, coloration and RMT lab reaction are the steps to detect the H1N1 virus. Pandemic Influenza Detection by Electrically Active Magnetic Nanoparticles and Surface Plasmon Resonance is a journal article by Kamikawa et al. (2012) [8]. stated that finding H1N1 by using various techniques, including the creation of nanoparticles, the use of glycans and polyanilines, and the modification of sensors, The crucial strain sequence utilized from NCBI and Sequence alignment which helps vaccine efficiency for influenza were discussed by Jerald et al. [9] in their research entitled "Influenza virus vaccine efficacy based on conserved sequence alignment." Research paper titled “Predicting H1N1 Vaccine Uptake and H1N1-Related Health Beliefs: The Role of Individual Difference in Consideration of Future Consequences” [10]. In this paper they attempt to assess the influence of individual difference in CFC on H1N1 vaccine uptake and get the result as participants who had talked with a health provider about the vaccine had a significantly higher likelihood of getting vaccinated. In their study titled "Aptamer-modified CNTFET biosensor for detecting H1N1 virus in a droplet," Huang et al. [11] proposed combining APTS and SAM immersed in a nanotube to produce CNTFET, which functions as a biosensor utilized to detect H1N1 virus by droplet. A descriptive study of vaccination coverage across hospitals and healthcare districts for three seasons was performed on the paper [12] to analyze the data on infection control indicators such as influenza vaccine coverage among healthcare workers (HCWs). In aspect of the researcher, vaccination coverage in Finland has been continuously high, most likely due to the present legislative framework. In their article titled "Signal-processing-based bioinformatics approach for the identification of influenza," Chrysostomou et al. (2013) [13] Neuraminidase genes were primarily discussed in "A virus subtypes in Neuraminidase genes,"F-score, Support Vector Machines, and signal processing are the techniques utilized to identify the influenza virus.

III. METHODOLOGY

In this paper, we examined data for predicting the probability of how likely one receives vaccine namely H1N1 vaccine and seasonal vaccine.

A. Preprocessing

(SET 1)-The training and testing dataset contains 26707 observations (or information taken in the survey) with 36 features obtained from the survey on the testing data. Out of these 36 features, the ’respondent_id’ column was dropped as it had no role in predicting the probability of vaccine and it is a unique number assigned to each person. Second, we check for the null values in the data, we took two approaches to deal with missing data in numerical columns, one by setting threshold as 18 and delete all the rows having null value greater than 18. Another we impute all the missing values by

Median of that particular column and categorical by assigning unknown to the missing values. There were no duplicates in the data Lastly, we need to convert all categorical values in dataset to numerical this was done using a python library called one hot encoding. We deleted the columns hhs_geo_region, employment industry by looking correlation graph. Resulting in SET 1 contains 51 columns in training set.

(SET 2)- As a second approach to deal with missing values we adopt the same method by dropping columns with setting threshold as 18, the used Simple Imputer Function with most frequent option to fill null values. Same method is used for both numerical and categorical variable. Using dummy variables, we encode all categorical variables using One hot encoder, here in the second approach we didn't drop any column before modelling. The total number of features after this is 106.

A. Data Analysis

Before building a machine learning model, we used graphs and tables to analyze datasets, understand trends, and determine the components for building a good machine learning model. This is part of a process aimed at gaining insight from historical data before making predictions about the future. Includes all techniques related to inspecting, cleaning, and detecting patterns in data.

1) Basic Information about dataset: This part consists of checking the datatypes, shape of the data, unique values in each column, null value using various functions like dtypes(), unique (), info (). Using describe () function provides a statistical insight into the data. This helps in identifying the measurements like mean, median, mode, interquartile range and wide variety of other information as well. This is the starting point for analyze various patterns in the data before proceeding to the next steps.

2) Pair Plot: A pair plot, is a visualization technique that can be used to explore relationships between multiple variables in a data set. A matrix of scatterplots is provided that plots each variable against all other variables in the dataset. Pair plots are useful for gaining insight into relationships and patterns within a data set, especially when examining interactions between multiple variables simultaneously. They help identify potential correlations, clusters, outliers, or trends in data.

3) Histograms: Since majority the features employed for predicting the probability of taking vaccine in this study are numeric, histograms complemented the information obtained from the describe table. The histograms represent the numerical data based on the bins assigned. Each break or bin is represented with a bar or frequency of the attribute. These graphs canbe tweaked to visualize additional details like the mean of a particular feature. We used hist plot to get histograms of all features.

4) Box-plots: These plots are another way of visualizing the skewness of numerical data. We used this technique to identify outliers in our dataset. It may be important to discard certain data points that introduce noise into the data.

5) Scatter-plots: In addition to analyzing each variable in the dataset separately, we also compare specific characteristics to see if there are trends between them. Scatter plots are useful when you want to show the relationship between two parameters. Use pair plots to get scatterplots of different features in your dataset.

5) Pivot Table: Pivot tables are a way of transforming and summarizing data in a tabular format. It allows you to group and aggregate data based on one or more variables. This tables provide a convenient way to summarize and restructure data, making it easier to perform data analysis and create visualizations to gain insight from your data.

6) Correlation Matrix: This is a graph that is used to determine the linear relationship between the different parameters in the dataset. The graph is sub-divided into different cells andeach represents the link between two attributes. Shows how a variable is correlated with each other variable. Value 1 indicate strong positive correlation, -1 indicate negative correlation and zero indicate no correlation between the variables.

7) Pie Charts: Pie charts are effective for visualizing any form of categorical data. As the analysis performed in this paper mainly seeks to predict the likelihood of two vaccine, this helped to visualize any changes in the ratio of the received vaccine rate. This is essential since it facilitates in the detection of any instances of skewed data.

B. Train Test Split

The data training-to-test split ratio affects model performance in machine learning. The proportions are randomly determined because the individual observations are independent of each other. We split the dataset into 70 parts.30 for modeling purposes. This means that 70 percent of the total data is used for model training and 30 percent is used as test data. Finally, the model's predicted observed values are compared to the target values in the test dataset.

E. Supervised Learning

In supervised learning we use labelled data to train the model to learn and identify patterns in the data and make predications from these models for unseen input data. Based on the training data, the algorithm modifies its internal parameters or weights to enhance prediction accuracy.

1)Logistic Regression: Logistic Regression is the one of the popular supervised learning algorithms which is best suited for binary classification. It models the relationship between dependent variable and one or more independent variable using the logistic function. It sets the threshold value as 0.5, if the probability obtained is below 0.5 then it is considered as 0 and if its above 0.5 then it is considered as 1. We implemented two separate logistic regression model one for h1n1 vaccine and other for seasonal vaccine with required parameters. We performed hyper tuning to get the best parameters.

1) Random Forests: These are decision trees that can be used for supervised machine learning. They can be used for both regression and classification. In this context we used this for classification. It combines individual decision tree results into a single result. The result from each decision tree is taken and the majority result is assigned as the final output [16]. Decision trees are often prone to overfitting to avoid these we performed cross validation with 5 folds.

1) *Support Vector Machine:* Support vectors machine are supervised machine learning models. This model can be used classify the data or even to predict the future values. SVM can be used to predict the probability of both vaccines. A virtual hyperplane created by SVM separates the data into various groups. Since it can resolve data into classes in higher dimensions as well, it is a potent method. [15]. The type of hyperplane to be implemented is chosen using the kernel function. This situation uses a rbf kernel function to separate the data. We did hyper tuning to get the optimal parameters for the model to improve the accuracy.

4)*Gradient Boosting Classifier:* This is one of the powerful supervised machine learning methods that combine multiple weak decision trees into a single one to create a strong predictive model. This model is based on the concept of boosting where each model is trained to correct the mistake of the previous one. We implemented two gradient boosting classifiers separately for predicting the probability of h1n1 and seasonal vaccine.

5)*Naïve Bayes Classifier:* Naïve bayes classifier is also a supervised learning method that is based on bayes theorem and it assumed that independence between the features. The Naive Bayes Classifier determines the prior probability, or the likelihood that each class will appear in the dataset, for each class. Additionally, it determines each feature's conditional probability, which expresses the possibility of detecting that feature given the class. Here we considered multinomial bayes classifier to create the model.

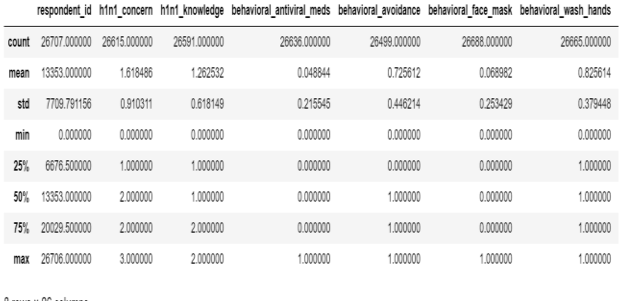
6)*Decision Tree Classifier:* A supervised learning technique called Decision Tree Classifier creates a tree-like representation of decisions and potential outcomes. Recursively dividing the data into subsets depending on the features allows the Decision Tree Classifier to create a tree structure. Based on a specific criterion, such as Gini impurity or information gain, the algorithm chooses the appropriate feature to split the data at each node. Maximizing class purity within each split is the objective. We implemented decision tree number of leaf as 10 and value of split equal to 8.

B. *Unsupervised Learning*

Unlike supervised learning, which uses labelled data to train the model, unsupervised machine learning does not. This method feeds the entire dataset into the model without separating it into training and testing sets. We specifically remove the tags from the target column in order to categories the observations. For cluster analysis, this type of machine learning model is frequently utilized. We did not use any kind of unsupervised learning model for this study.

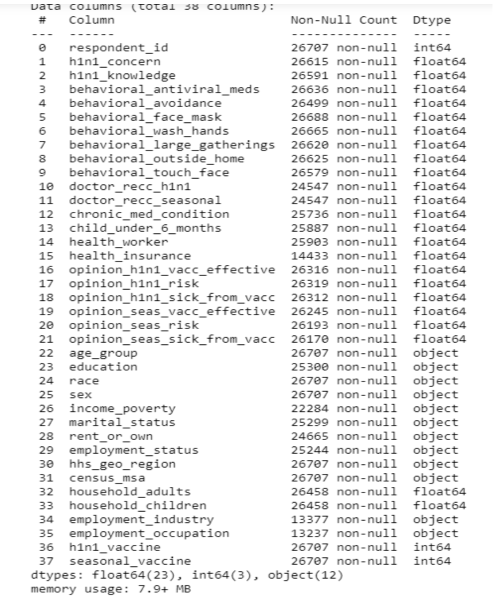
IV. RESULTS

We performed analysis and predicted the likelihood of one receiving the H1N1 and seasonal vaccine. As part of the preliminary analysis, we examined two types of statistics tables. The descriptive statistics table, as seen in Figs. 1 and 2, provides a more comprehensive view into the dataset by providing information on datatypes and missing values, mean, median, quartile range and many other aspects.



	respondent_id	h1n1_concern	h1n1_knowledge	behavioral_antiviral_meds	behavioral_avoidance	behavioral_face_mask	behavioral_wash_hands
count	26707.000000	26615.000000	26591.000000	26636.000000	26499.000000	26688.000000	26665.000000
mean	13353.000000	1.618486	1.262532	0.048844	0.725612	0.068982	0.825614
std	7709.791156	0.910311	0.618149	0.215545	0.446214	0.253429	0.379448
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	6676.500000	1.000000	1.000000	0.000000	0.000000	0.000000	1.000000
50%	13353.000000	2.000000	1.000000	0.000000	1.000000	0.000000	1.000000
75%	20029.500000	2.000000	2.000000	0.000000	1.000000	0.000000	1.000000
max	26706.000000	3.000000	2.000000	1.000000	1.000000	1.000000	1.000000

Fig. 1. Descriptive Statistics Table for a part of the Dataset



```
Data columns (total 38 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   respondent_id                26707 non-null  int64
 1   h1n1_concern                 26615 non-null  float64
 2   h1n1_knowledge               26591 non-null  float64
 3   behavioral_antiviral_meds    26636 non-null  float64
 4   behavioral_avoidance         26499 non-null  float64
 5   behavioral_face_mask         26688 non-null  float64
 6   behavioral_wash_hands        26665 non-null  float64
 7   behavioral_large_gatherings  26620 non-null  float64
 8   behavioral_outside_home      26625 non-null  float64
 9   behavioral_touch_face        26579 non-null  float64
 10  doctor_recc_h1n1             24547 non-null  float64
 11  doctor_recc_seasonal         24547 non-null  float64
 12  chronic_med_condition        25736 non-null  float64
 13  child_under_6_months         25887 non-null  float64
 14  health_worker                25903 non-null  float64
 15  health_insurance             14433 non-null  float64
 16  opinion_h1n1_vacc_effective  26316 non-null  float64
 17  opinion_h1n1_risk            26319 non-null  float64
 18  opinion_h1n1_sick_from_vacc  26312 non-null  float64
 19  opinion_seas_vacc_effective  26245 non-null  float64
 20  opinion_seas_risk            26193 non-null  float64
 21  opinion_seas_sick_from_vacc  26170 non-null  float64
 22  age_group                    26707 non-null  object
 23  education                    25300 non-null  object
 24  race                         26707 non-null  object
 25  sex                          26707 non-null  object
 26  income_poverty               22284 non-null  object
 27  marital_status               25299 non-null  object
 28  rent_or_own                  24665 non-null  object
 29  employment_status            25244 non-null  object
 30  hhs_geo_region               26707 non-null  object
 31  census_msa                   26707 non-null  object
 32  household_adults             26458 non-null  float64
 33  household_children           26458 non-null  float64
 34  employment_industry          13377 non-null  object
 35  employment_occupation        13237 non-null  object
 36  h1n1_vaccine                 26707 non-null  int64
 37  seasonal_vaccine             26707 non-null  int64
dtypes: float64(23), int64(3), object(12)
memory usage: 7.9+ MB
```

Fig. 2. Information Table for a part of the Dataset

To visualize the number of people taken H1N1 vaccine or not (Fig. 3) and number of people taken seasonal vaccine and not taken vaccine (Fig.4), we used count-plots and pie charts (fig.5) to get an idea about the proportion of each vaccine. We discovered the data was slightlyskewed to the seasonal vaccine; majority of the people are likely to receive seasonal vaccine where 75% people are less like to take h1n1 vaccine.

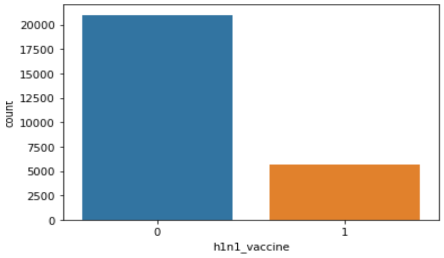


Fig.3. Count plot of h1n1 vaccine

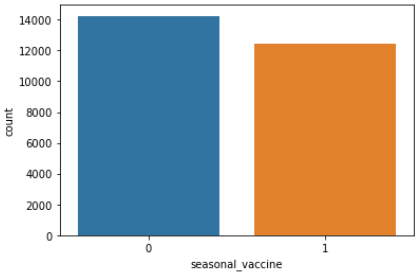


Fig.4. Count plot of seasonal vaccine

From the count plot it is clear that people are more likely to take seasonal vaccine than h1n1 vaccine.

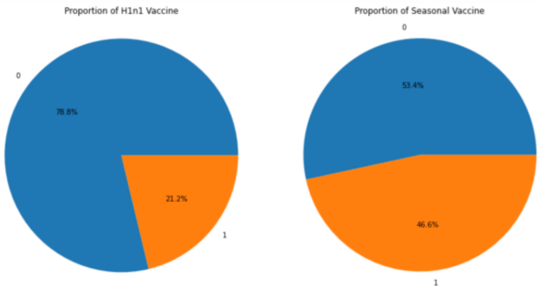


Fig.5. Pie chart

Numerous graphs, such the histogram (Fig. 6) and boxplot (Fig. 7), reveal information about certain numerical parameters. The histograms can be used to identify outliers in the data, but the box plots make them stand out more. Here we plotted between age group and h1n1 vaccine and there are no outliers. The hist plot gives the count of various values present in each feature.

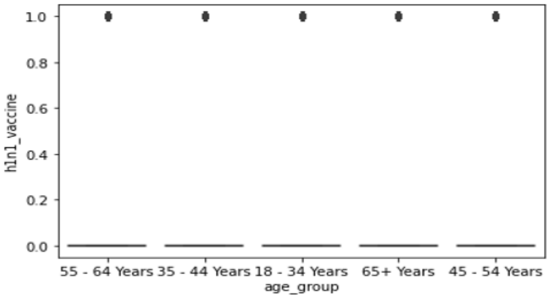


Fig.7. Box plot of age group vs h1n1 vaccine

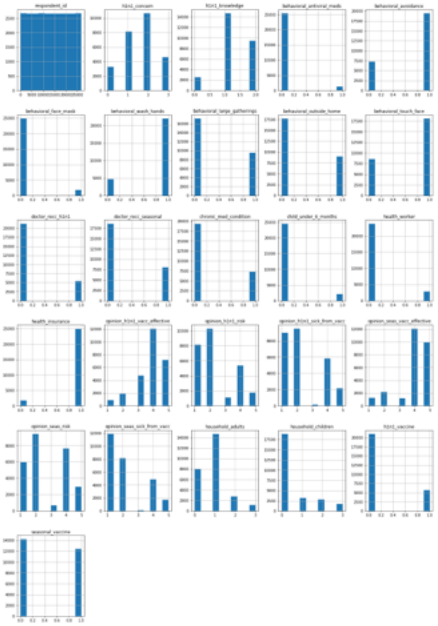


Fig.6. Histogram plot

The pair-plots helped in finding a link between any two features of the dataset. As seen from Fig. 8, we can find a relationship between all features to another features.

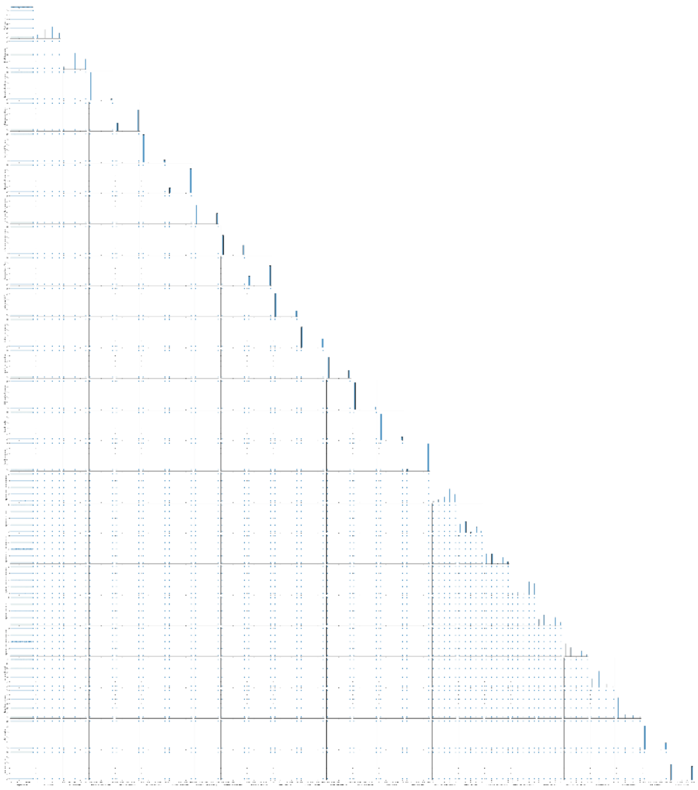


Fig.8. pair plot of the dataset

We also created pivot tables of some features and compare it with seasonal and h1n1 vaccine so we get better comparison of features. In this we compared various features like age group (Fig.9.), education (Fig.10), and income status (Fig.11)

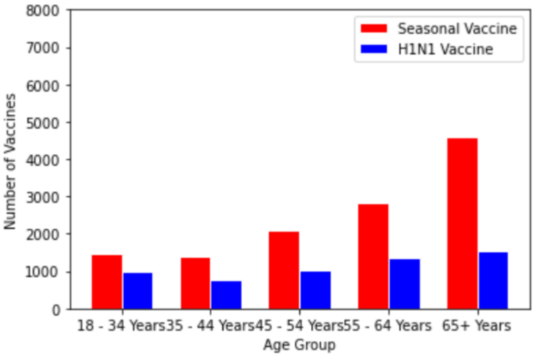


Fig.9. Vaccine by Age group

From this plot it is clear that aged person was took vaccine than the other age groups. Among the all-age group the seasonal was the one received most. By comparing the vaccine by education, it is come to know that graduates were received more vaccine than any other educational level with seasonal vaccine preferred.

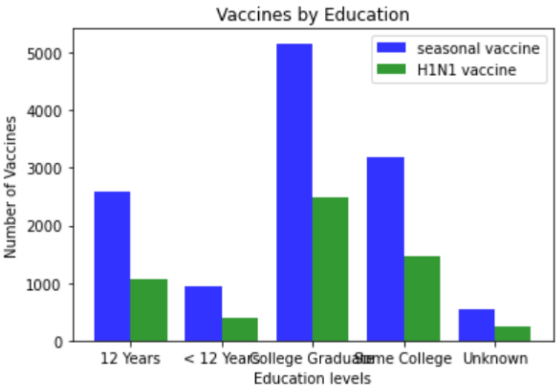


Fig.10. Vaccines by education

While comparing the vaccine by income group there was an remarkable trend that those who are belong to below poverty were received more vaccine.

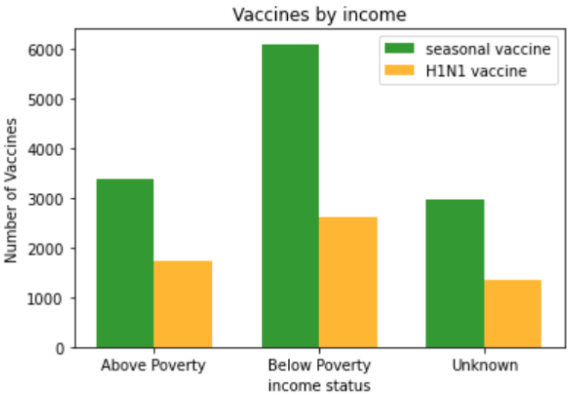
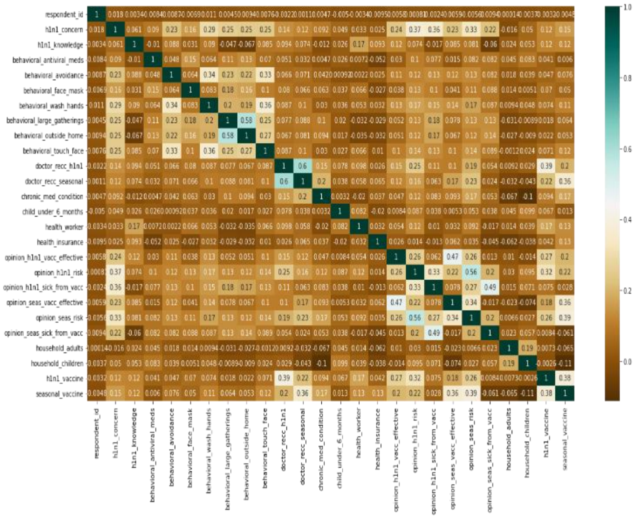


Fig.11. Plot for vaccine by income



We plotted the correlation (Fig between the features given from this plot we get the how each variable is correlate with another variable and this helps us to delete 2 features by looking at the correlation value. All these features correlation coefficient was almost zero so we eliminate them from modelling.

To reduce the number the features to see if it had a better impact on the machine learning model predictions, we performed a principal component analysis [14]. The plot below (Fig.12) gives the cumulative variance explained by number of principal components.

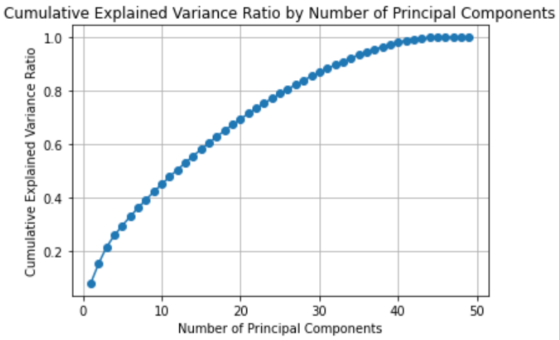


Fig.12. Principal Component Analysis

A confusion matrix was used by the majority of the machine learning models in this study to cross-validate the projected values. The confusion matrix was used to calculate the accuracy, precision, and recall. (Fig .13) shows the confusion matrix obtained in random forest classifier for h1n1 vaccine and (Fig.14) depict the confusion matrix obtained for the seasonal vaccine using random forest classifier.

We plot ROC (Receiver Operating Characteristic) Curve (Fig.15) for the model as a measure of evaluation. Each point on the curve represents a different threshold setting, and the curve provides a visualization of how the model's performance changes as the threshold is adjusted.

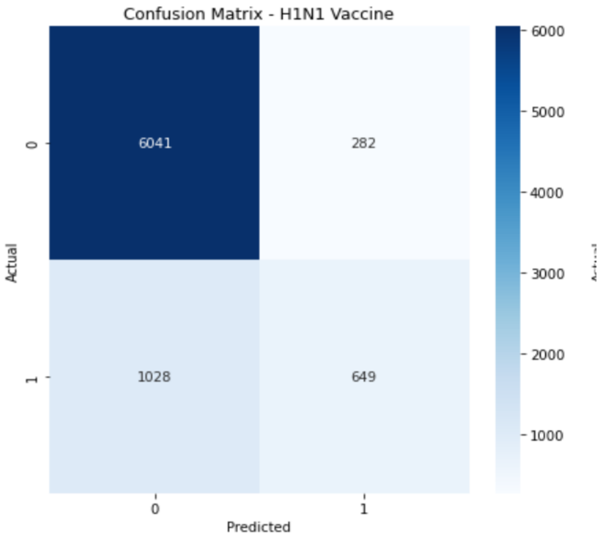


Fig.13.Confusion matrix of H1N1 vaccine using random forest

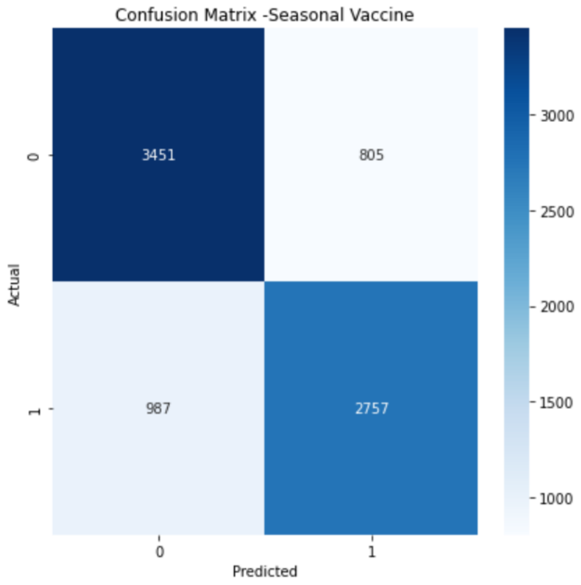


Fig.14.Confusion matrix of Seasonal vaccine using random forest

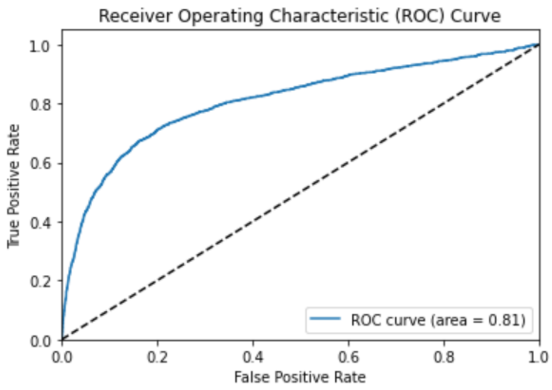


Fig.16. ROC Curve

IV. DISCUSSION

As seen from the above figures it is clear that in our dataset people are more likely to take seasonal vaccine and less likely to take h1n1 vaccine. The majority of the people took seasonal vaccine on the other side vast majority reject to take the h1n1 vaccine. We know that there are people who would like to receive both vaccine in concern with their health. Since the majority of data was given in binary values than continuous numerical values there was no presence of outliers, we checked this with the boxplot. And there is no need to normalize the data as well. We eliminate the null values at a threshold of 18 and rest of the null values we impute with median for numerical columns and unknown value for categorical columns. In order to fit the data set into a machine learning model we change all the categorical values to numerical using one hot encoding.

The relationship between the characteristics in our dataset could not be effectively visualized using a correlation table, it was discovered, this is because we employed a large data set with an excessive number of features. A correlation matrix cannot use it because of the enormous number of columns. But we eliminate 4 features as by looking the different values of that features we found that is irrelevant for modelling.

Compared to pie charts, count plots are more useful for effectively visualizing the frequency of a particular cancer cell type. This is as a result of pie charts showing show proportions as angles. The human eye has trouble comprehending the essence of anything until the proportionate difference is substantial. A bar plot will therefore aid in understanding the situation when visualizing datasets with a small proportional difference.

We use python’s describe function to describe the features in different parameter terms in the dataset. We could see the count, minimum value, maximum value, mean, median, quartile ranges and standard deviation. Since feature values follow binary format most of the parameters for categorical columns are almost similar only variation is for standard deviation. Using python info () function we can find the null values and data types.

Histograms also assisted in finding the count of values in each feature list. The pair plot draws scatter plot for two different features, since the number of features in the dataset are high, we can’t insight anything from these, as it looks congested. The plot using pivot table helps us to compare feature with two different vaccines at the same time. From these plots we can see that in all the compared features seasonal vaccine was the one predominating the h1n1 vaccine. People are less likely to receive h1n1 vaccine. Old aged people and graduate are the one category who took vaccine more than other value

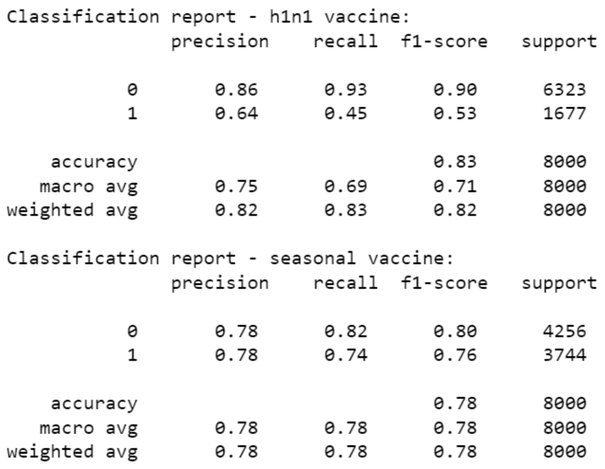
We then split the dataset in the ratio of 70:30 for training and testing respectively. That is 70 % data is used for training the various machine learning model and 20 % of the data is used for testing purpose. We Set the value of random state as 42 this ensures the reproducibility of the dataset.

When we compared different machine learning model in this data, we found that most of the supervised machine learning methods we implemented performs in a good manner. We didn’t implement any kind of unsupervised machine learning techniques for this dataset. We implemented logistic regression, Random Forest classifier, Support vector machine, Naïve bayes classifier and Decision Tree classifier to predict how likely a person receives the h1n1 and seasonal vaccine. For each of this model we implemented two instances model separately for h1n1 vaccine and seasonal vaccine. We fitted the model using training data and then predict the outcome using the created model for testing data.

Overall, the support vector machine is best and most successful in predicting the probability of how likely a person receives the h1n1 vaccine and gradient boosting classifier for seasonal vaccine with given features. The SVM model has an accuracy of 84 percentage for H1N1 vaccine which highest of all other models. So SVM performs better for H1N1 vaccine. The SVM was initially trained with linear kernel then later we applied hyper tuning and select the best parameters for the fitted model.

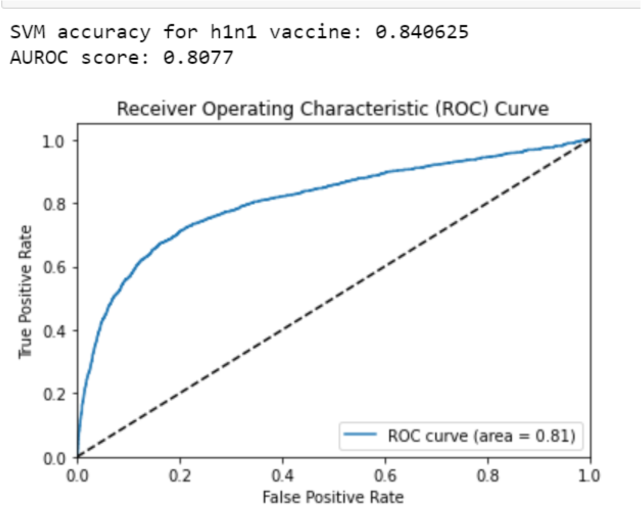
The gradient decent classifier is found to be better in predicting seasonal vaccine with an accuracy of 74. We hyper-tune this model with parameters to get the best result. The forest model performs the second for both H1N1 and seasonal vaccine with an accuracy of 83.62 percent for H1N1 and 77.76 percent for seasonal vaccine. Then comes the logistic regression model with probability prediction accuracy of 83 percent and 77 percent accuracy for H1N1 and seasonal vaccine. Decision tree also performs in an average manner for predicting the probability of vaccine with 82 for H1N1 and 61 for seasonal vaccine. Among the all, the least performance was done by naïve bayes classifier for both H1N1 and seasonal vaccine the accuracy is around 77 percentage only.

For testing the accuracy of different models, parameters like precision, recall, classification report, confusion matrix, ROC curve was used. Both accuracy and recall must be examined in order to evaluate an algorithm's performance fully. Precision is defined as the proportion of relevant occurrences among all retrieved instances. Recall, also known as sensitivity, is the fraction of recovered occurrences among all relevant cases. We are more interested in the recall of predicting probability of vaccine. In our SVM model, the recall for vaccines is 93 percent value 0 and 45 percent for 1 for h1n1 vaccine. For seasonal vaccine the recall is 82 and 74 for 0 and 1 respectively. We also look into the AUROC score for all the models. The classification report for the SVM model is given in (fig 16) By computing the area under this curve, the AUROC score provides a summary of the model's performance across all feasible criteria. An increase in the score, which runs from 0 to 1, denotes superior performance. The obtained AUROC score for the SVM model is 0.84 for H1N1. The ROC curve for the SVM model can be seen in fig (17.a,17. b).



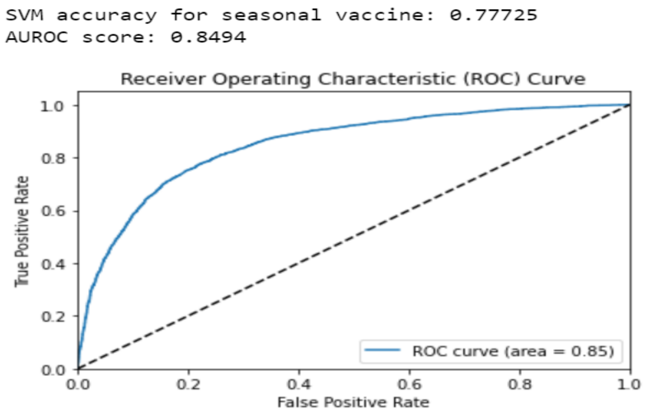
```
Classification report - h1n1 vaccine:
              precision    recall  f1-score   support

           0       0.86      0.93      0.90      6323
           1       0.64      0.45      0.53      1677

    accuracy                           0.83      8000
   macro avg       0.75      0.69      0.71      8000
weighted avg       0.82      0.83      0.82      8000

Classification report - seasonal vaccine:
              precision    recall  f1-score   support

           0       0.78      0.82      0.80      4256
           1       0.78      0.74      0.76      3744

    accuracy                           0.78      8000
   macro avg       0.78      0.78      0.78      8000
weighted avg       0.78      0.78      0.78      8000
```

Fig 16: Classification report for SVM



```
SVM accuracy for h1n1 vaccine: 0.840625
AUROC score: 0.8077
```

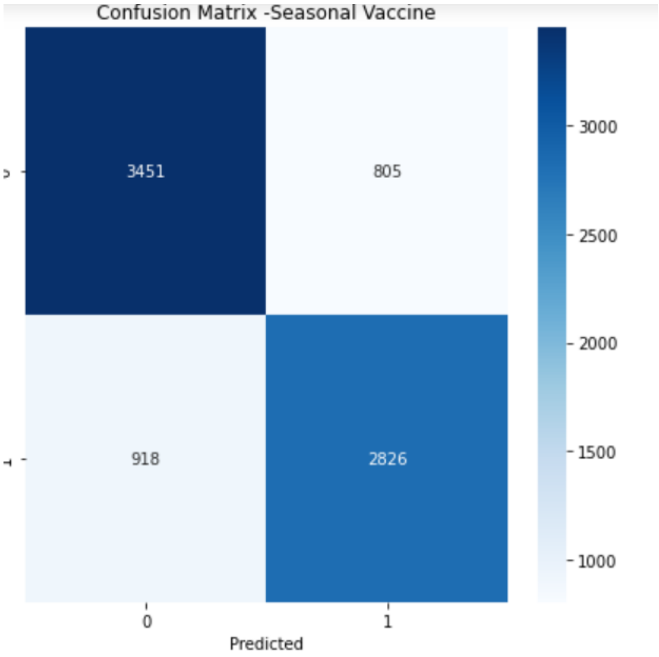
Fig 17.a ROC curve of H1N1 vaccine

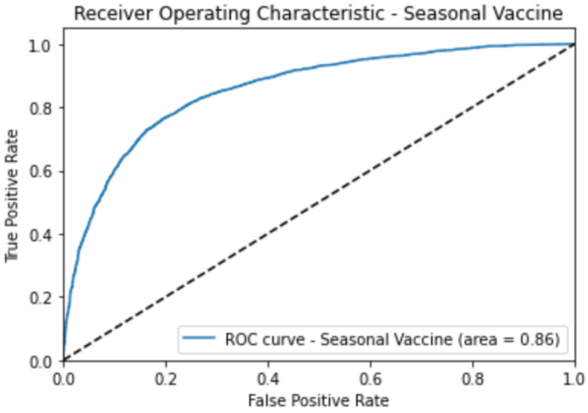


```
SVM accuracy for seasonal vaccine: 0.77725
AUROC score: 0.8494
```

Fig 17.a ROC curve of Seasonal vaccine

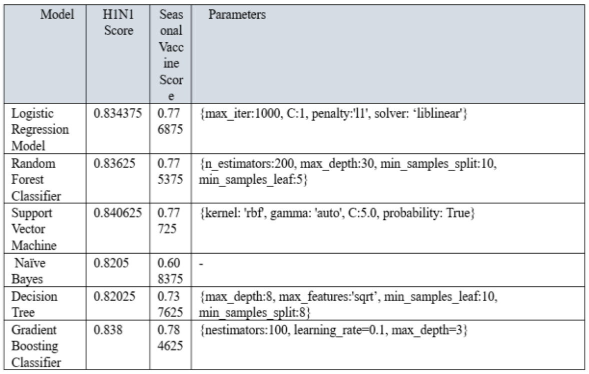
The confusion matrix and ROC Curve for the Gradient Boosting classifier for seasonal vaccine is given in Fig (18).





The ROC Score obtained for the seasonal vaccine using ss gradient boosting is 83, which is a good score.

The Table 1 shows the comparison of all the implemented machine learning model from SET 1 (Best of two)



Model	H1N1 Score	Seasonal Vaccine Score	Parameters
Logistic Regression Model	0.834375	0.776875	{max_iter:1000, C:1, penalty:'l1', solver: ‘liblinear'}
Random Forest Classifier	0.83625	0.775375	{n_estimators:200, max_depth:30, min_samples_split:10, min_samples_leaf:5}
Support Vector Machine	0.840625	0.77725	{kernel: 'rbf', gamma: 'auto', C:5.0, probability: True}
Naïve Bayes	0.8205	0.608375	-
Decision Tree	0.82025	0.737625	{max_depth:8, max_features:'sqrt’, min_samples_leaf:10, min_samples_split:8}
Gradient Boosting Classifier	0.838	0.784625	{nestimators:100, learning_rate=0.1, max_depth=3}

Table 1: Models and Score

COMPARISON OF MODELS

Starting with Logistic regression which is one the best classifier that can be used for binary classification. Table 2 provides the comparison of result obtained using 2 different datasets. From the table it is clear that Logistic regression with SET 1 gave more accuracy for both vaccines compared to SET 2. We did hyper tuning with SET 1 to obtain the best parameters that fit the model.

Set	H1N1 Score	Seasonal Score
SET 1	0.8343	0.7768
SET 2	0.82283	0.7581

Table 2. Scores of Logistic Regression

The second model implemented is Random Forest classifier with both datasets SET 1 and SET 2. In this machine learning model, we got almost same accuracy as logistic regression with SET 1 and for SET 2 82.87 for H1N1 and 75.64 for seasonal vaccine. In this model SET 1 performs better than SET 2. Table 3 shows the scores of the model. We selected the parameters of model using hyper tuning and cross validation with SET 1.

Set	H1N1 Score	Seasonal Score
SET 1	0.8362	0.7753
SET 2	0.8287	0.7564

Table 3. Scores of Random Forest

Table4. depicts results of Support Vector machine (SVM), which is identified as the best model among all for predicting seasonal vaccine. We obtained an accuracy score of 84 % for h1n1 vaccine and 77% for seasonal vaccine. The AUROC score obtained is 80 and 84 with SET 1. Using the SET 2, the score obtained is 78 and 53 for H1N1 and seasonal vaccine respectively. Performed cross validation with SET 1.

Set	H1N1 Score	Seasonal Score
SET 1	0.8465	0.7855
SET 2	0.777	0.5337

Table 4. Scores of SVM

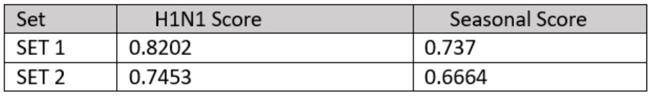
Naïve Bayes classifier modelled for predicting the probability of vaccine gives lowest accuracy score for seasonal vaccine with a score of 60 % with SET 1 but SET 2 gives more accuracy for seasonal vaccine at a rate of 72%. The score obtained for h1n1 vaccine is 78 and 82% for SET 2 and SET 1 respectively.

Set	H1N1 Score	Seasonal Score
SET 1	0.8205	0.6083
SET 2	0.7825	0.7244

Table 5. Scores of Naïve Bayes Classifier

Another model implemented using SET 1 and SET 2 is decision tree classifier. For SET 1 we did hyper-tuning and obtained optimal parameters where in SET 2 we modelled it without hyper-tuning the model. The Accuracy obtained for SET 1 is higher than the SET 2. Table 6 compare the model scores.

We did hyper tuning to the model to improve its efficiency with SET 1. Table 6 shows the different scores of the model.



Set	H1N1 Score	Seasonal Score
SET 1	0.8202	0.737
SET 2	0.7453	0.6664

Table 6: Scores of Decision Tree

VI.CONCLUSION

Based on data from the Nation H1N1 Flu Survey "2009" (NHFS) and the Centers for Disease Control (CDC), this study attempts to predict both H1N1 and seasonal flu vaccination rates. Initially, we performed some analysis to understand the attributes associated with data. We built 6 supervised machine learning model to predict the outcome. The dataset underwent techniques called imputation and one hot encoding while preprocessing. We did hyper tuning for models to find the best parameters that fit the data. H1N1 vaccination prediction is performed best by an SVM model with an RBF kernel and hyperparameter tuning using GridSearchCV, which produced an accuracy of 83.43%, while seasonal flu vaccination prediction is performed best by gradient boosting classifier, which produced an accuracy of 78.46%. This vaccination aids in preventing both the seasonal flu and the H1N1 virus. In order to provide the public with the necessary information regarding the significance of the H1N1 vaccine and seasonal flu vaccine in 2009, awareness was raised throughout all social media platforms. It was noted that the younger demographic was more severely impacted than the population over 65. Vaccinations were made available to immunize the population and to provide a safe environment for everyone.

VII. FUTURE RESEARCH SCOPE

Despite the positive results, this study has serious shortcomings. Twitter is generally not used regularly for temporal and spatial data collection. This data discrepancy can change and adversely affect model performance. Differences in accuracy exist between regional and national levels that can be affected by lack of accurate data, as individuals from the same region often exhibit similar behavioral characteristics. In the future, improvements in technology will increase the quality and quantity of data, which may lead to improved performance and problem analysis. Further information about seasons, especially non-pandemic seasons, would be very useful for research in this project. In the coming time we looking forward to implement more machine learning algorithms to obtain better result.

REFERENCES

[1] 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. 10.1109/CCWC.2019.8666474.

[2] T. T. S. Chinh, D. H. Stephanus, C. Kwoh, C. Schönbach and X. Li, "A possible mutation that enables H1N1 influenza a virus to escape antibody recognition," 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Hong Kong, 2010, pp. 81-84, doi: 10.1109/BIBM.2010.5706541.

[3] Bao Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. The Influenza Virus Resource at the National Center for Biotechnology Information. J. Virol. 2008 Jan; 82(2): 596-601.

[4] Wei Hu, “Molecular features of highly pathogenic Avian and Human H5N1 Influenza a virus in Asia," Comput. Mol. Biosci., vol. 2, no. 2, pp. 45–59, 2012.

[5] F. Stalder and J. Hirsh, “Open source intelligence,” First Monday, vol. 7, no. 6, 2002. 416.

[6] N. Wiriyachaiporn, H. Sirikett and T. Dharakul, "Rapid influenza an antigen detection using carbon nanostrings as label for lateral flow immunochromatographic assay," 2013 13th IEEE International Conference on Nanotechnology (IEEE-NANO 2013), Beijing, 2013, pp.166-169,doi:10.1109/NANO.2013.6720979.

[7] Y. Ma, W. Chang, C. Wang, H. Ma, P. Huang and G. Lee, "An integrated passive microfluidic device for rapid detection of influenza a (H1N1) virus by reverse transcription loop-mediated isothermal amplification (RT-LAMP)," 2017 19th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS), Kaohsiung, 2017, pp. 722-725, doi:10.1109/TRANSDUCERS.2017.7994150.

[8] Kamikawa, Tracy & Norton (Mikolajczyk), Malgorzata & Kennedy, Michael & Zhong, Lilin & Zhang, Pei & Setterington, Emma & Scott, Dorothy & Alocilja, Evangelyn. (2012). Pandemic Influenza Detection by Electrically Active Magnetic Nanoparticles and Surface Plasmon Resonance. Nanotechnology, IEEE Transactions on. 11. 88 - 96. 10.1109/TNANO.2011.2157936.

[9] A. Baby Jerald and T. R. Gopalakrishnan Nair, "Influenza virus vaccine efficacy based on conserved sequence alignment," 2012 International Conference on Biomedical Engineering (ICoBE), Penang, 2012, pp. 327-329, doi: 10.1109/ICoBE.2012.6179031.

[10] 1.Xiaoli Nan & Jarim Kim (2014) Predicting H1N1 Vaccine Uptake and H1N1-Related Health Beliefs: The Role of Individual Difference in Consideration of Future Consequences, Journal of Health Communication, 19:3, 376-388, DOI: 10.1080/10810730.2013.821552

[11] J. Huang, T. Lin, W. Chang and W. Hsieh, "Aptamer-modified CNTFET biosensor for detecting H1N1 virus in droplet," The 4th IEEE International NanoElectronics Conference, Tao-Yuan, 2011, pp. 1-2, doi: 10.1109/INEC.2011.5991640.

[12] Hammer Charlotte C, Lyytikäinen Outi, Arifulla Dinah, Toura Saija, Nohynek Hanna. High influenza vaccination coverage among healthcare workers in acute care hospitals in Finland, seasons 2017/18, 2018/19 and 2019/20.Euro Surveill. 2022;27(17): pii=2100411.https://doi.org/10.2807/15607917.ES.2022.27.17.2100411.

[13] Chrysostomou, C., & Seker, H. (2013). Signal-processing-based bioinformatics approach for the identification of influenza A virus subtypes in Neuraminidase genes. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 3066-3069.

[14] Abdi, H., Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459. https://doi.org/10.1002/wics.101.

[15] Suthaharan, S. (2016). Support Vector Machine. Machine Learning Models and Algorithms for Big Data Classification, 207–235. https://doi.org/10.1007/978-1-4899-7641-3-9.

[16] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha and S. Kundu," Improved Random Forest for Classification," in IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 4012-4024, Aug. 2018, doi: 10.1109/TIP.2018.2834830.