Predicting If A Population Have Received Their H1N1 And Season Flu Vaccines Using The 2009 FLU Survey

Sahil Rai
School of Computer Science
The University of Nottingham
Nottingham. United Kingdom
efysr3@nottingham.ac.uk

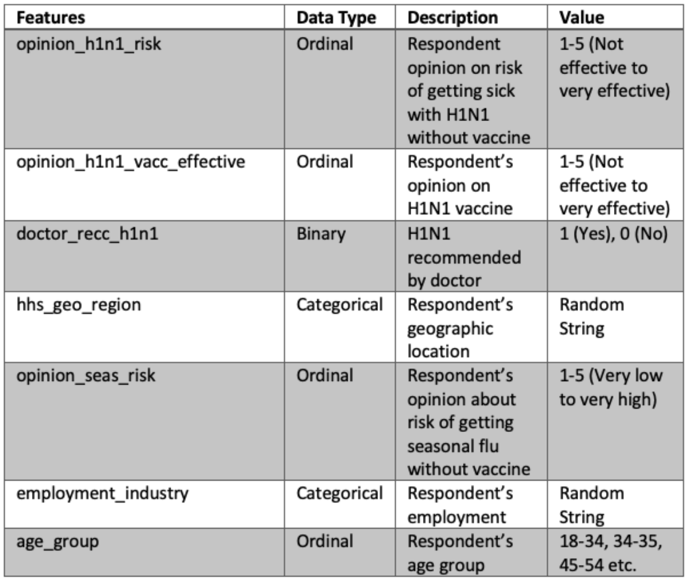
Deonte Allen-Gooden
School of Computer Science
The University of Nottingham
Nottingham. United Kingdom
efyda1@nottingham.ac.uk

***Abstract*—This paper investigates if it is possible to predict whether the target population of 6 months or older living in the United States received their H1N1 and seasonal flu vaccines. This is based on investigating shared information on backgrounds, opinions, and health behavior. The data used in this study is from the National 2009 H1N1 Flu Survey [1] which will undergo various data analysis practices. Sections of the paper will include the hypothesis of which practice to use, evidence of data extrapolation, and evaluation of the data modeling techniques. Furthermore, the paper will contain visualisation techniques to help illustrate the vigorous data analysis methods utilised. The paper aims to provide research on the following main sections, Research Question, Literature review, Methodology, Discussion, Conclusion, and future recommendations.**

***Index Terms*—H1N1 vaccines, seasonal flu vaccines, National 2009 H1N1 Flu Survey, KNN, Decision Model, SVM, Random Forest**

I. Introduction

H1N1 and seasonal flu are both types of influenza viruses but are slightly different. H1N1 is a strain of type A influenza that caused the 2009 pandemic, known as Swine Flu. In response to stop the spread of swine flu, the H1N1 vaccine was developed which is effective for just that specific strain. On the other hand, seasonal flu is a combination of type A and B influenza. The seasonal flu is common and expected to circulate every year therefore seasonal vaccines are improved annually [2]. Both viruses have similar symptoms, however, H1N1 can be more severe than the common flu. There are always personal characteristics on reasons why an individual would get vaccinated with prejudice against or for vaccines [3]. Our initial hypothesis lies based on this prejudice, and we believe that individuals will have higher seasonal flu vaccination rates in comparison to H1N1 vaccination because of how well-informed people are om the seasonal flu. However, there are many various reasons to consider in order to make a clear and accurate judgment, therefore using the National 2009 H1N1 Flu Survey, we will study individual behaviors, geographic, and social characteristics to extrapolate trends. This will help us answer the question "Is it possible to predict if a target population has received their H1N1 and seasonal flu vaccines".



Features	Data Type	Description	Value
opinion_h1n1_risk	Ordinal	Respondent opinion on risk of getting sick with H1N1 without vaccine	1-5 (Not effective to very effective)
opinion_h1n1_vacc_effective	Ordinal	Respondent's opinion on H1N1 vaccine	1-5 (Not effective to very effective)
doctor_recc_h1n1	Binary	H1N1 recommended by doctor	1 (Yes), 0 (No)
hhs_geo_region	Categorical	Respondent's geographic location	Random String
opinion_seas_risk	Ordinal	Respondent's opinion about risk of getting seasonal flu without vaccine	1-5 (Very low to very high)
employment_industry	Categorical	Respondent's employment	Random String
age_group	Ordinal	Respondent's age group	18-34, 34-35, 45-54 etc.

Fig. 1. Data types from the National 2009 H1N1 Flu Survey

A. Data Set

This section describes the initial findings from the data set. The dataset holds responses from 26,707 individuals and has 35 unique responses each. The survey shows diversity in data types which can be seen in Fig 1.

B. Pre-process and data-wrangling

Prior to any analysis or visualisation to help identify our prediction. The data must first be pre-processed including various data-wrangling tactics to further improve the data in terms of usability and accuracy. Upon first data inspection, the data set will need to undergo data integration with the labels for improved readability, which can then be classed as our main data frame. After this, a data cleaning can proceed which will include identifying errors, inconsistencies, or missing data. An initial look gives us the indication that there is missing data that will need to be filled or dropped based on a threshold that we decide on. We can also further the data reduction, by using methods such as dropping unnecessary columns of data that will not aid our conclusion and by identifying any extreme cases or outliers however, this dataset may be hard to find those due to the constraints of the variables. Feature engineering is something we could specifically use from our data an example of this could be the combination of our target variables 'h1n1_vaccine' and 'seasonal_vaccine' to create a combined target variable which would aid the visualisation element as well as enable us to answer further questions such as the probability of people having both vaccines. Finally, we could do some data exploration including visualising our data so that we can identify any patterns, trends, and relationships helping to guide our analysis and identify if there are any issues with the data incredibly early on.

II. Literature Review

The medical industry has been transformed due to the powers of machine learning. Machine learning allows for efficient classification and predictions for multiple diseases. Reading several papers surrounding our topic we reviewed the methods of existing papers using the same H1N1 dataset and other papers that used different datasets.

G.G. Giambrone et al. [4] "Influenza Vaccination and Respiratory Virus Interference Among Department of Defense Personnel During the 2017-2018 Influenza Season" predicts the effectiveness of the H1N1 vaccine against the influenza virus, using combined datasets. Giambrone used vaccine coverage and calculation of person-time at risk to help efficiently predict the likeliness of catching the influenza virus. Furthermore, this methodology led to the findings of 47% effectiveness for the vaccine. The findings of this study predicted that vaccination was associated with reduced risks of influenza viruses. The training of data and the calculation of time risk allows for more efficient conclusions. Despite this, we find studies that use other methodologies for results deemed more desirable.

Nemesure, M.D. et al "Predictive modeling of depression and anxiety using electronic health records and A Novel Machine Learning Approach with artificial intelligence"[5] predicts the risk of depression and anxiety based on patient electronic health records (EHR). The paper explores different machine learning algorithms such as Support Vector Machines (SVM), Convolution Neural Networks (CNN), and Recurrent Neural Networks (RNN). This methodology of training a deep neural network on the EHR data set reported predicting risks with high accuracy, and in their evaluation, it was better than other machine learning methods. Overall, the paper concludes the important significance of machine learning techniques in the medical field.

Dwyer et al "Machine Learning Approaches for Clinical Psychology"[6], explores questions associated with treatment prediction, diagnosis, prognosis, and detection, using Machine learning methodologies with priority to develop an accurate model. The method mentioned walks through each step of; Pre-processing which includes scaling, imputation; Hyper-parameter optimisation for the best performance scores; and algorithm selection based on effectiveness. The paper concluded the Regularisation hyper-parameter was essential to provide a good accuracy score and SVM was effective for this problem with a predictive accuracy of above 70% exceeding the accuracy threshold. Overall, the paper concludes the use of SVM provided a balanced predictive accuracy score and that research in clinical psychology and psychiatry can be highly benefited by applying machine learning.

D. Lavanya et al "Ensemble Decision Tree Classifier for Breast Cancer Data" [7]. Highlighted the use of data mining, pre-processing, and feature selection in their methodology. Furthermore, with the medical diagnosis, this was seen as highly important. They highlight the use of multiple decision tree algorithms within the paper adopting the significance of the CART (Classification and Regression Trees) as a powerful method for classification especially for medical data. Analysing different variations of the CART method, using hybrid methods or feature selection to improve the all-around accuracy of the classification. During this, the presented hybrid approach which includes the feature selection method, bagging, and the cart method gave the most optimal results with an accuracy of 74.47%, 97.85%, and 95.86%. However, these high results are extremely close to CART with feature selection leaving these methods as the most optimal when using a decision tree method. In our paper, we aim to have optimal accuracy when obtaining our results.

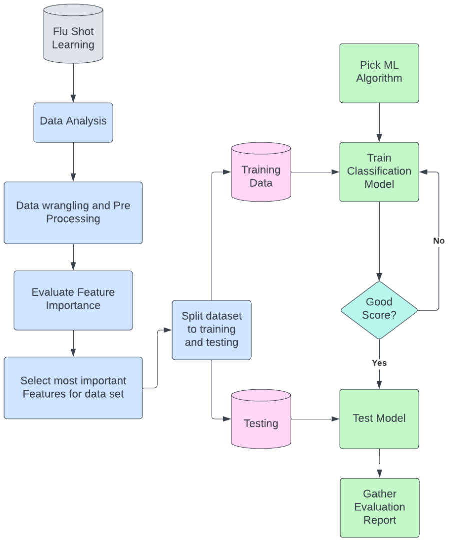

Flu Shot Learning
Data Analysis
Data wrangling and Pre Processing
Evaluate Feature Importance
Select most important Features for data set
Split dataset to training and testing
Training Data
Testing
Pick ML Algorithm
Train Classification Model
No
Good Score?
Yes
Test Model
Gather Evaluation Report

Fig. 2. Machine Learning process

III. Methodology

Our study will be using five different types of supervised machine learning classification models (K Nearest Neighbor, Decision Tree Classifier, Naive Bayes, Kernal SVM, and

Random Forest) this is so we can produce a comparative study. Each classification model will have a contrast in pre-processing steps and analysis, producing varied accuracies. Fig 2. is a flow chart illustrating the general methods for each model.

A. Author Contribution

In this paper, Method 1 is explored by Deonte Allen-Gooden, and Method 2 is explored by Sahil Rai respectively. The remainder of the paper is written collaboratively.

B. Data Exploration and Analysis

To begin the steps taken to explore this binary classification problem, it is imperative to first have a thorough look at the data. We used the ".info()" python method which provides a summary of the data set. This allowed us to identify data types and counts of non-missing values. The Vaccine data set had some object types and nulls value which incurred, hence a form of encoding is necessary to transform the object data types to have a numeric type which is essential for machine learning computation. We will need to explore processes to handle the missing data, as these discrepancies can cause inaccuracies and biases in the model. We decided to take a different pre-processing approach which can be seen in sections C and D, however, we decided to use the top 7 features from our Random Forest Feature Importance classifier seen in Fig 3.

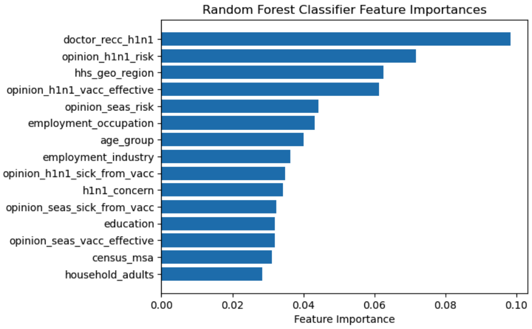


Fig. 3. Feature Selection Importance

In this data set our target values are Yes (1) and No (0) respective to having taken H1N1 and seasonal vaccines. The binary values can be seen as classes, and a simple bar graph can be applied to visualise the class distribution as seen in Fig 4. There is a balanced class distribution for seasonal vaccines however there is a class imbalance in response to H1N1 vaccines. This may cause a bias in the model, and therefore further methods such as resampling, class weighting, or using specialised algorithms can be considered to allow a more balanced training environment.

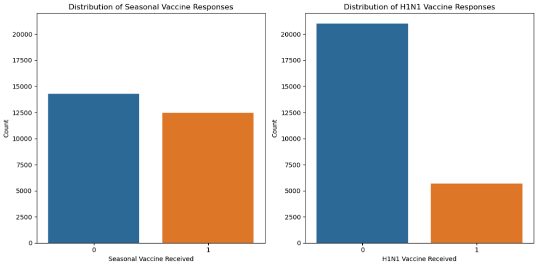


Fig. 4. Class Distribution

To look into relationships across the features, there are different methods such as using bar graphs, density curves, and box plots. Fig 5. shows count distribution of each selected feature using box plots. Immediately we noticed that there is a perfect discriminator, where the response to doctor_recc_h1n1 feature shows a 100% count rate for yes (1). This model, emphasizes the class imbalance mentioned earlier but also justifies the use of feature selection since a perfect discriminator feature is considered of high importance for machine learning models [9]. Conclusively, we decided to not resample the data and we elaborate further in the Discussion Section.

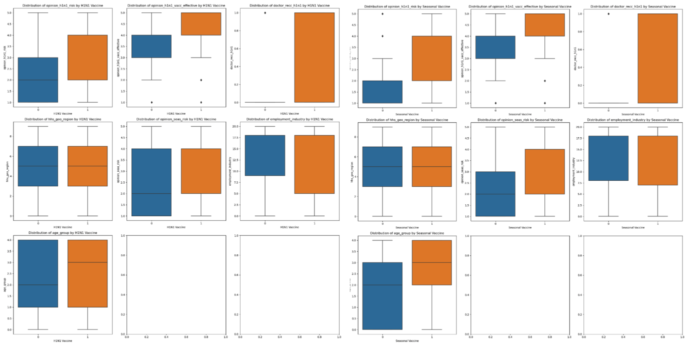


Fig. 5. H1N1 and Seasonal vaccine feature trends using Box plots

C. Method 1 - Decision Tree and KNN Classification Model

1) Pre-processing: I made use of the whole train and test datasets for my models. The general clean-up of these data sets included the analysis of data types and the entry of any missing data points. The strategy used for this was the use of the Simple Imputer with the "most frequent" occurring strategy to fill in all the missing data. Furthermore, I then encoded the data into categories using the Ordinal Encoder giving all my non-binary data numerical groups. Fig 6. is the formulae for the standardisation that I used, I scaled this data using the Standard Scaler feature to bring values as close to 0 or 1 as possible. After this was complete, I deemed my data fit for training.

$$z = \frac{x - \mu}{\sigma}$$

Fig. 6. Standard Normalisation Formulae

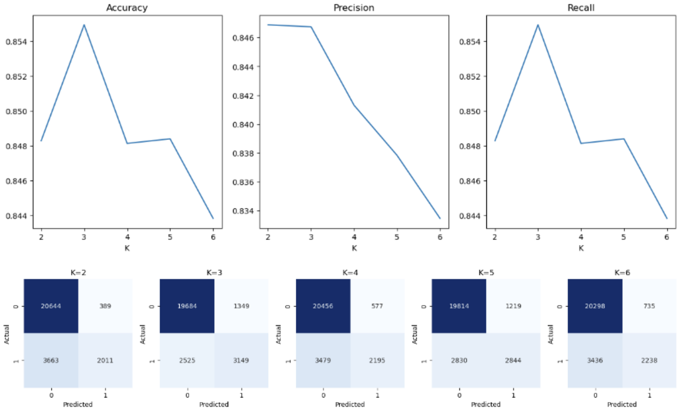


Fig. 7. Evaluation of optimal K value

2) Data Classification: For classification, I decided to use and analyse the results of three different classifiers so I could pick the top two. I used KNN, Naive Bayes, and the Decision Tree Classification Models. To begin I started by defining my train and test data sets. Due to the robustness of KNN and Naive Bayes models and their non-parametric nature, I opted to use these algorithms which can evaluate linear and non-linear data which is great for our data which is mostly binary. Using the Decision Tree classifier, I had a random state of 42 to run multiple instances of this classifier for constant behavior and results. In terms of optimal parameters for Naive Bayes, there are no parameters to tune, in general, Naive Bayes is not affected by the curse of dimensionality [15] Decision Tree classification has parameters to tune, despite this, I did not choose to deviate from the default parameters due to the results given. Furthermore, the use of Decision Trees will excel in our condition because of the class distribution, and Decision Trees are also the easiest for implementation [15]. K-nearest neighbors I decided to evaluate the parameters that I used in terms of K. Using a loop of K values, I visualised the K values for optimal results against accuracy, precision, and recall. This can be shown in Fig 7. which is plots and confusion matrices of my K values. Illustrated here having a K value of 3 gives me an optimal return. The use of KNN classification is good as it does not require any training and helps include any abnormal occurrences[15].

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

Fig. 8. Naive Bayes Classifier

D. Method 2 - SVM and Random Forest Classification Model

1) Pre-processing: In the pre-processing stage after data analysis, the problem at hand was the presence of missing data and object data types. To tackle the missing data problem I chose to use random imputation to randomly fill in missing values. Random imputation handles missing data especially well as the responses have no continuous relation and each property is distributed on ”random” responses. This technique mitigates bias in comparison to other techniques such as mean, mode, or regression imputation, and sought the same benefits of handling missing data. I also noticed the presence of categorical variables. Machine learning algorithms require numeric values in order to compute models, and this encoding step is essential, especially for SVM and RF models. Therefore, I opted to use label encoding which transforms categorical variables into numeric variables. Label Encoding allowed me to keep the ordinal structure where some features had an ordered structure in response i.e. Age, ranges from youngest to oldest, and preserves the dimensionality by keeping numeric variables in the same column. I chose this method in comparison to one-hot encoding which is susceptible to the ”curse of dimensionality” which has its own issues like overfitting and computational complexity.

2) Data Classification: RF and SVM are both supervised machine learning models, taking in independent (Selected Features) which are associated with dependent target variables (Binary Response).

SVM is a discriminative classifier that uses supervised learning to produce the most optimal hyperplane [11]. SVM utilises parameter tuning, which includes regularisation parameters, gama, and kernel. To find the most optimal hyperparameter for SVM classifies, I applied grid search using cross-validation. The search grid Cross-validation technique resamples the training data into k-folds in which I opted for a standard utilisation of 3-folds, this then filters through all potential SVM classifiers giving a more accurate configuration. This automated process recommends the most optimal parameters and can be tuned accordingly.

On the other hand, a Random Forest classifier uses a set of decision trees that tally each prediction based on uncorrelated models to create a predictive product by reviewing the tally [13]. This fits with our data as it poses predictive power, previously mentioned in the data analysis section, and is uncorrelated. I then utilised a random search cross-validation technique for hyper tuning which is similar to the search grid cross-validation technique but reduced computational complexity.

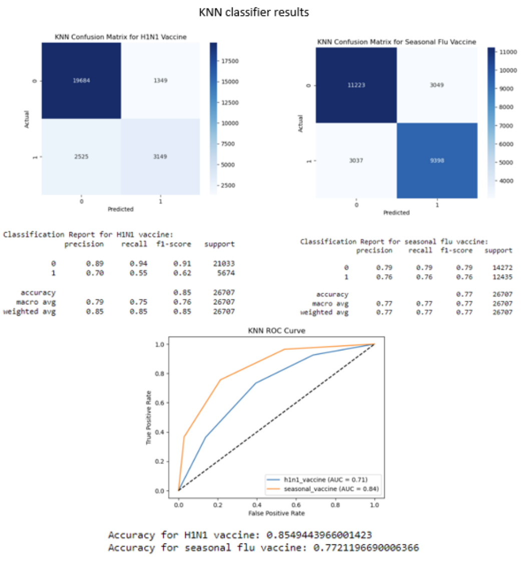
IV. Results

In this section, we will present the results of our four models through visualisation by presenting the accuracy and the ROC (receiver operating characteristic) curves of our chosen models. Furthermore, through evaluation, we can highlight a desired model for final processing.

We also performed another data visualisation of our selected features after standardisation seen in in Fig 11. This allows us to evaluate further our data after distributional changes due to scaling. This visualisation can be later be used to understand deviations in the data set.

A. Method 1 - Decision Tree and KNN Classification Results

Through the process of systematical elimination, I have chosen only two models to take forward for results and evaluation and this is the Decision Tree and KNN classifiers due to higher accuracy in results.



```
Classification Report for H1N1 vaccine:
              precision    recall  f1-score   support

           0       0.89      0.94      0.91     21033
           1       0.70      0.55      0.62      5674

    accuracy                           0.85     26707
   macro avg       0.79      0.75      0.76     26707
weighted avg       0.85      0.85      0.85     26707
```

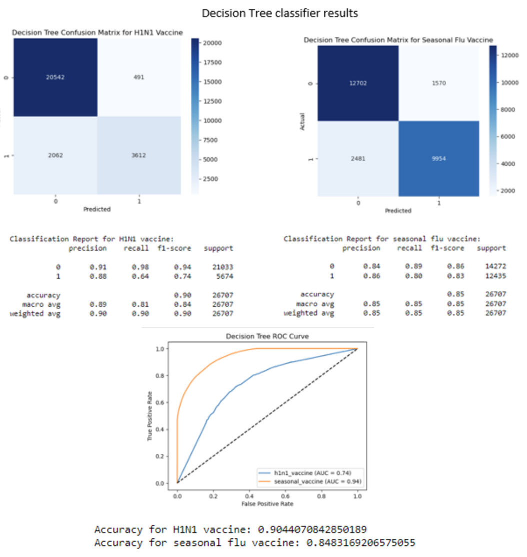
```
Classification Report for seasonal flu vaccine:
              precision    recall  f1-score   support

           0       0.79      0.79      0.79     14272
           1       0.76      0.76      0.76     12435

    accuracy                           0.77     26707
   macro avg       0.77      0.77      0.77     26707
weighted avg       0.77      0.77      0.77     26707
```

```
Accuracy for H1N1 vaccine: 0.8549443966001423
Accuracy for seasonal flu vaccine: 0.7721196690006366
```

Fig. 9. KNN results



```
Classification Report for H1N1 vaccine:
              precision    recall  f1-score   support

           0       0.91      0.98      0.94     21033
           1       0.88      0.64      0.74      5674

    accuracy                           0.90     26707
   macro avg       0.89      0.81      0.84     26707
weighted avg       0.90      0.90      0.90     26707
```

```
Classification Report for seasonal flu vaccine:
              precision    recall  f1-score   support

           0       0.84      0.89      0.86     14272
           1       0.86      0.80      0.83     12435

    accuracy                           0.85     26707
   macro avg       0.85      0.85      0.85     26707
weighted avg       0.85      0.85      0.85     26707
```

```
Accuracy for H1N1 vaccine: 0.9044070842850189
Accuracy for seasonal flu vaccine: 0.8483169206575055
```

Fig. 10. Decision Tree Results

The KNN classifier scored an accuracy of 85% for the H1N1 vaccine and scored 77% for the seasonal flu vaccine. Whereas, for my Decision Tree classifier we had an accuracy of 90% for the H1N1 vaccine and 85% for the seasonal flu vaccine. Using the accuracy as one of the main metrics for evaluation we see that the decision tree classifier has a considerably higher accuracy. Using the classification report function I also have the precision, recall and f1-scores for each model. For each classifier I have an AUC score which is a more comprehensive overview of the data and does not get affected by the class imbalance, the AUC score of the KNN classifier for H1N1 was 71% and seasonal flu of 84%. Result of the AUC for the Decision Tree classifier was 74% for H1N1 and 94% for seasonal.

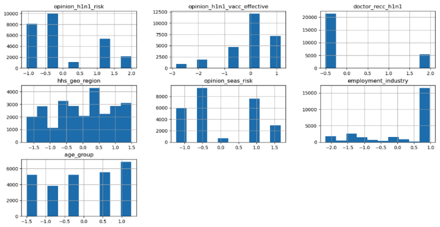
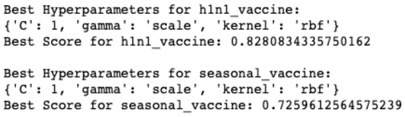


Fig. 11. Feature Distribution after Standarisation

B. Method 2 - SVM and Random Forest Classification Results

To get the best predictive results each step of data analysis, pre-processing, and classification must be carefully selected. In the pre-processing step, I applied two types of hyperparameter tuning, RF utilising Random search using cross-validation and SVM using grid search with cross-validation. In this sub-section, I will evaluate the results derived from hyperparameter tuning.

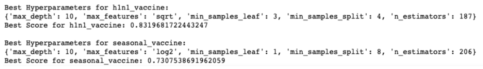


```
Best Hyperparameters for h1n1_vaccine:
{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}
Best Score for h1n1_vaccine: 0.8280834335750162

Best Hyperparameters for seasonal_vaccine:
{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}
Best Score for seasonal_vaccine: 0.7259612564575239
```

Fig. 12. Hyper Parameter Tuning of SVM classifiers

Fig 12. shows the parameter that was used to tune the SVM model which concurs the following: Regularisation held a lower value which allows the model fitting to be worked more flexibly, Gamma is set based on scaled based on the data set which sources for the most optimum attributes and uses a Radial Basis Function (rbf) Kernel which is a non-linear kernel that computes non-linear relation of target and feature variables[12]. This tuning aims to provide a more accurate model.

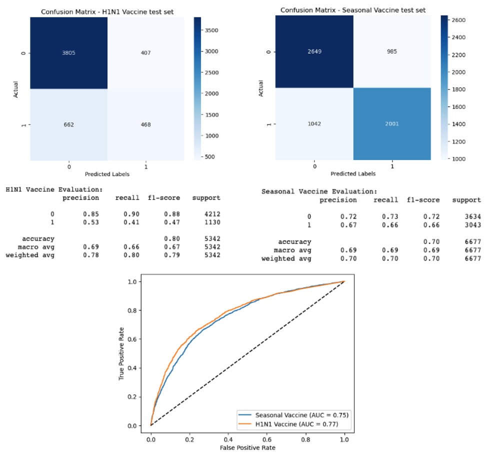


```
Best Hyperparameters for h1n1_vaccine:
{'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 3, 'min_samples_split': 4, 'n_estimators': 187}
Best Score for h1n1_vaccine: 0.8319681722443247

Best Hyperparameters for seasonal_vaccine:
{'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 8, 'n_estimators': 206}
Best Score for seasonal_vaccine: 0.7307538691962059
```

Fig. 13. Hyper Parameter Tuning of RF classifiers

Fig 13. shows the most optimal parameter configuration for the RF classifier. This includes ’max_depth: higher value can take in more complexity, ’max_features’: increase data area and randomness, ’min_samples_leaf’: minimum number of lead node values, ’min_samples_split’: minimum split of internal node, ’n_estimators’: number of decision trees to evaluate and gives the best score for both h1n1_vaccine and seasonal_ vaccine. These parameters tweaking can lead to underfitting or overfitting of data in the model and the search grid aims to find the perfect balance to produce the most accurate model whilst leveling computational complexity at a reasonable state. To evaluate my metrics I utilised a combination of, classification_report (Python scikit-learn function) consisting of precision, recall & f1-score, confusion matrix, and a Reciever Operating Characteristic (ROC) curve. These can be seen in Fig. 14. and Fig 15. respectively for both SVM and RF models.



```
H1N1 Vaccine Evaluation:
              precision    recall  f1-score   support

           0       0.85      0.90      0.88      4212
           1       0.53      0.41      0.47      1130

    accuracy                           0.80      5342
   macro avg       0.69      0.66      0.67      5342
weighted avg       0.78      0.80      0.79      5342
```

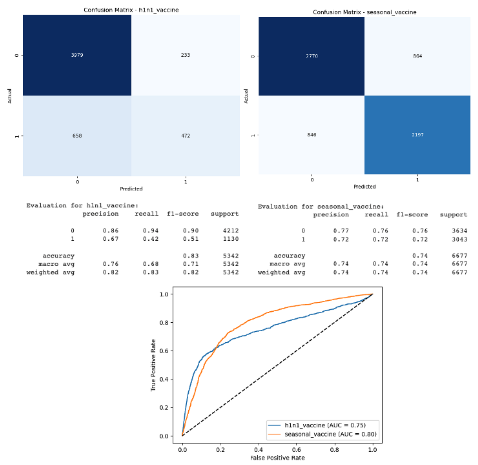
```
Seasonal Vaccine Evaluation:
              precision    recall  f1-score   support

           0       0.72      0.73      0.72      3634
           1       0.67      0.66      0.66      3043

    accuracy                           0.70      6677
   macro avg       0.69      0.69      0.69      6677
weighted avg       0.70      0.70      0.70      6677
```

Fig. 14. Random Forest results

The random forest model produced an accuracy of 80% for H1N1 vaccine prediction and a 70% accuracy score for the seasonal vaccine. Furthermore, boasting 79% weighted average on f1_score for H1N1. Whilst scoring 70% for the seasonal vaccines. F1- scores weigh both precisions and recall therefore giving a good judgment on the performance of the classification model, therefore it suggests that Random forest performed better in predicting H1N1 predictions. Furthermore, H1N1 vaccine produced an AUC of 77%, and the seasonal vaccine produce an AUC of 75% which provides a performance indicator and is unaffected by class imbalance giving a more comprehensive indication of performance for our data set.



```
Evaluation for h1n1_vaccine:
              precision    recall  f1-score   support

           0       0.86      0.94      0.90      4212
           1       0.67      0.42      0.51      1130

    accuracy                           0.83      5342
   macro avg       0.76      0.68      0.71      5342
weighted avg       0.82      0.83      0.82      5342
```

```
Evaluation for seasonal_vaccine:
              precision    recall  f1-score   support

           0       0.77      0.76      0.76      3634
           1       0.72      0.72      0.72      3043

    accuracy                           0.74      6677
   macro avg       0.74      0.74      0.74      6677
weighted avg       0.74      0.74      0.74      6677
```

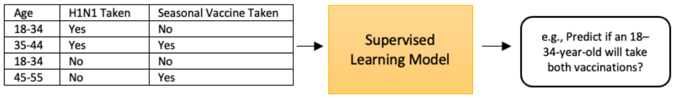
Fig. 15. SVM Results

Similarly, the SVM model (Fig 15.) produced accuracy levels of above 70% with an accuracy of 81% for H1N1 vaccines and 74% for seasonal vaccines.

V. Discussion

The two approaches that we took had different methods at various stages of analysis this is starting from the pre-processing method down to the final results. We used a range of techniques that we deemed appropriate for our data set and based our models on that thought process.

Firstly data analysis for both methods started by focusing on the data spread from the H1N1 National survey[1]. The vaccination data set consists of labels taking in independent variables from our Selected Features (e.g. Age) which are associated with our dependent target variables of binary responses (Yes or No to having taken vaccination). We concluded to categorise this as a supervised learning problem and the model can be seen in Fig 16. Supervised learning gives us the chance to have a clear understanding of labeled categorised training data. There are Studies such as ”Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation” by M. Morid that make use of supervised learning for prediction [13]. On the other hand, Unsupervised learning would be more suitable where pattern recognition and relationships between clusters are the main focus. Further to this, we sub-categorised from the choices of either regression models or classification models, due to the discrete properties in our variables we opted to pick classification models, specifically we picked KNN, Naive Baye, Decision Tree, and Random Forest for our first approach and Support Vector Machines models for our second approach.



Age	H1N1 Taken	Seasonal Vaccine Taken
18-34	Yes	No
35-44	Yes	Yes
18-34	No	No
45-55	No	Yes

Supervised Learning Model
e.g., Predict if an 18–34-year-old will take both vaccinations?

Fig. 16. Example of a Simplistic Supervised Learning Model

Further to our data analysis, we employed data visualisation techniques to assess the distribution of features and classes. But this comes with the hurdle of handling high-dimensional as our initial data set consisted of 35 features. To overcome this hurdle, we opted to use feature selection, which will evaluate the most effective features. This approach derives further benefits like visualising a subset of data, reducing overfitting, and improving model performances. Ngyuen C. et al ”Random Forest classifier combined with feature selection for breast cancer diagnosis and prognosis” [8] focuses on the important step of feature selection to identify the most relevant features to their classification problem. The authors utilised feature selection with a random forest (RF) classifier which derived a predictive accuracy of around 99.8%. This prompted us to go forward with a similar approach and utilise the RF classifier alongside feature selection to provide the best features for our model, this results can be seen in Fig 3.

In Fig 4. our class distribution is presented, and for H1N1 we found a class imbalance which can lead to bias in some models. Further analysis from Fig 5. showed us a box plot distribution and the doc_recc_h1n1 feature showed all responses on one of the classes; which is know as the ”perfect discriminator”. The book ”Programs for Machine Learning”. Morgan Kaufmann [9] mentions the impact of a perfect discriminator, specifically in association with a Decision Tree model. This poses decision tree handling, with a perfect discriminator providing a clear decision rule where an ideal binary test set could lead to a perfect classification. Through this, we can pose the question ”Does having a perfect discriminator feature lead to better classifications of the Decision Tree model compared to its counterparts?”. Per this question, we decided to not re-sample the data set and keep the feature balance as it is as we were curious about the performance out come of using a Decision Tree Model compared to the other classification methods.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Fig. 17. Min-Max Formulae

Another hurdle was to choose between different scaling techniques. We chose the scaling technique of standard normalization formulae Fig 6. over min max Fig 17. Standardisation scales all features to a similar scale, this is beneficial as it removes bias from larger scales which in turn we hope for a fairer comparison of our predictive models [10]. This is an important step, for algorithms such as SVM and KNN model as they are distance-based algorithms, and the balanced scaling mitigates one variable over-weighting others.

Model 1 for processing made use of the whole test and train data sets and this helps to encourage an unbiased performance if the whole data set is considered. Furthermore, this reduces the workload for hyperparameter tuning in method 1. The ordinal encoding used in method 1 is used to categorise all non-numerical data and was deemed an easier method to group the unknown features such as employment occupation. Model 2 on the other hand made use of a 75-25 split for training and test sets with the SVM AND Random Forrest(RF) classifiers.

Method 1 had the model with the highest accuracy with a score of 90% for the H1N1 vaccine from the Decision Tree model this also had 85% accuracy for the seasonal vaccine Method 1 also has an H1N1 score of 85% and a seasonal score of 77% for the KNN classifier. Method 2 had scores of 80% and 70% for the H1N1 and seasonal flu vaccine respectively for the RF model. Furthermore, this method achieved a score of 81% and 74% for H1N1 and seasonal vaccines. Drifting onto the AUC score for the two highest scorers of each method the Decision Tree classifier had 74% and 94% for the H1N1 and seasonal flu vaccine whereas the SVM model has percentages of 75% and 80%. Therefore overall we can conclude that the highest in terms of accuracy and AUC score was the Decision Tree classifier with the default parameters. This is backed by the book ”Programs for Machine Learning.” by Morgan Kaufmann [9]. Fig 18. provides a graphical representation showing the most accurate models moreover it is to be noted that without the AUC scores for reference, we can see that all of the scores had higher accuracy in the H1N1 section this could be down to the features selection and could indicate the need for a class re-sample in the future so that both instances perform equally.

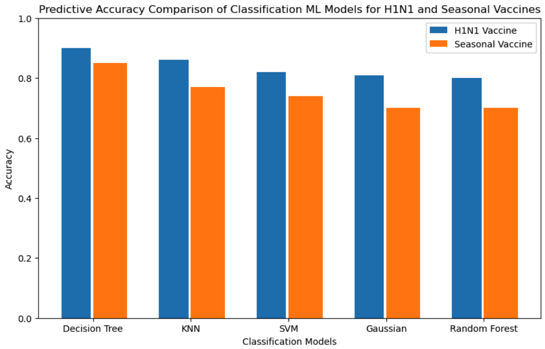


Fig. 18. Accuracy Comparison

Covering previous studies on the models we choose we can see that Z. Li et al [17] make use of an extremely sophisticated SVM method with high levels of parameter tuning looking at biased and unbiased samples the accuracy of this SVM was 96.9% for bias and 99.4% for unbiased samples demonstrating the high potential for this model. Despite these results, the reason we did not manage to achieve such accuracy for our model is that we opted for standardised methods, this is an example but an unfair comparison due the difference in data which causes a change in methods at all stages of analysis. A similar comparison would be a comparison of our Decision Tree model where the study ”Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms” by M. Tu et al [18] is a study that makes use of a decision tree in order to classify health issues within this study we have similar goals of classifying health probabilities on target variables and also the results for this study found a precision of 78% which is down to the data having less classification bias than we currently have. Furthermore, this method also does not analyse the parameters efficiently to have the highest possible accuracy. Finally, looking at a study that uses the same data as our study S. Inampudi et al [19] worked on a paper called ”Machine learning based prediction of H1N1 and seasonal flu vaccination” they used similar methods to analyse the prediction of the H1N1 and seasonal flu vaccine. Specifically within method 2 where both the study and our study make use of the Random Forrest model results in the study [19] achieve 82% and 85% for H1N1 and seasonal predictions whereas we have 83% and 73% receptively. This illustrates our H1N1 predictions are above average in accuracy for Random Forrest however our seasonal flu vaccine is below average which is down to a difference of feature selection. This evaluation of different methods and scores promotes for a change of features for a balanced score in accuracy.

VI. Conclusions

All of our models performed considerably well on our dataset with all results achieving an accuracy of over 70%. Upon inspection, we see that H1N1 accuracies all performed higher than the seasonal flu accuracies this could be due to the imbalances of our data. Despite, this our main question of whether we are able to predict if our target population had received their H1N1 or seasonal flu vaccine has been achieved through multiple models. Furthermore, our most successful model was the Decision Tree classifier in method 1 which has an accuracy of 90% for the H1N1 vaccine predictions. Nevertheless, future work would consist of eliminating class imbalances to have a greater representation of data.

Method 1 had the highest values out of both methods but if we were to retest parameters for the KNN and Decision Tree model by doing an increased hyperparameter tuning process we may have found higher accuracies than our results seen in Fig 10. Moreover, the K in our model is at optimal results for this model therefore this evaluation was done to a good standard for the KNN classifier.

Another improvement to our methodology which could impact and improve our results would be, to use hyperparameter tuning techniques to optimise the f1-score. A method we could use to do this would be to vary the threshold from values between 0 and 1 and see how this threshold affects the f1-score having a starting point of 0.5 and then consider classes as positive or negative if they are above or below this starting point. Studies such as ”Anomaly Detection: How to Artificially Increase your F1-Score with a Biased Evaluation Protocol” By D. Fourure et al [16] have shown methods on how to increase f1-scores through threshold manipulation.

References

[1] DrivenData (no date) Flu shot learning: Predict H1N1 and seasonal flu vaccines, DrivenData. Available at: https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/ (Accessed: March 2, 2023).

[2] ilani, T.N., Jamil, R.T. and Siddiqui, A.H. (2022) H1N1 influenza - statpearls - NCBI bookshelf. Available at: https://www.ncbi.nlm.nih.gov/books/NBK513241/ (Accessed: March 6, 2023).

[3] Murphy, J. et al. (2021) Psychological characteristics associated with covid-19 vaccine hesitancy and resistance in Ireland and the United Kingdom, Nature News. Nature Publishing Group. Available at: https://www.nature.com/articles/s41467-020-20226-9 (Accessed: March 6, 2023)

[4] G. G. Giambrone et al., ”Influenza Vaccination and Respiratory Virus Interference Among Department of Defense Personnel During the 2017-2018 Influenza Season,” Vaccine, vol. 37, no. 43, pp. 6527-6531, 2019.

[5] Nemesure, M.D. et al. (2021) Predictive modeling of depression and anxiety using electronic health records and A Novel Machine Learning Approach with artificial intelligence, Nature News. Available at: https://www.nature.com/articles/s41598-021-81368-4 (Accessed: 01 May 2023)

[6] Dwyer, D.B., Falkai, P. and Koutsouleris, N. (2018) Machine learning approaches for clinical psychology and psychiatry, Machine Learning Approaches for Clinical Psychology and Psychiatry. Available at: https://www.annualreviews.org/doi/full/10.1146/annurev-clinpsy-032816-045037 (Accessed: 01 May 2023).

[7] Lavanya, D. and Rani, K. (2012) Ensemble decision tree classifier for breast cancer data, International Journal of Advanced Information Technology. Available at: https://scholar.archive.org/work/vnw43d3xh5g35bjnzlxv36froy (Accessed: 01 May 2023).

[8] Nguyen, C., Wang, Y. and Nguyen, H.N. (2013) Random Forest classifier combined with feature selection for breast cancer diagnosis and prognostic, SCIRP Open Access. Available at: https://www.scirp.org/html/6-9101686_31887.htm (Accessed: 16 May 2023).

[9] Quinian, R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann.

[10] Liu, C. (2022) Data transformation: Standardization vs normalization, KDnuggets. Available at: https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html (Accessed: 08 May 2023).

[11] Patel, S. (2017) Chapter 2: SVM (Support Vector Machine) - theory, Medium. Available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 (Accessed: 17 May 2023).

[12] Kumar, S. (2021) 20x times faster grid search cross-validation, Medium. Available at: https://towardsdatascience.com/20x-times-faster-grid-search-cross-validation-19ef01409b7c (Accessed: 09 May 2023).

[13] Morid MA;Kawamoto K;Ault T;Dorius J;Abdelrahman S; (no date) Supervised learning methods for predicting healthcare costs: Systematic Literature Review and empirical evaluation, AMIA ... Annual Symposium proceedings. AMIA Symposium. Available at: https://pubmed.ncbi.nlm.nih.gov/29854200/ (Accessed: 01 May 2023).

[14] Yiu, T. (2021) Understanding random forest, Medium. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (Accessed: 09 May 2023).

[15] Glen, S. (2019) Comparing classifiers: Decision trees, K-NN &amp; Naive Bayes, Data Science Central. Available at: https://www.datasciencecentral.com/comparing-classifiers-decision-trees-knn-naive-bayes/ (Accessed: 01 May 2023).

[16] Fourure, D. et al. (2021) Anomaly detection: How to artificially increase your F1-score with a biased evaluation protocol, arXiv.org. Available at: https://arxiv.org/abs/2106.16020 (Accessed: 01 May 2023).

[17] Z. Li, R. Yuan and X. Guan, ”Accurate Classification of the Internet Traffic Based on the SVM Method,” 2007 IEEE International Conference on Communications, Glasgow, UK, 2007, pp. 1373-1378, doi: 10.1109/ICC.2007.231.

[18] M. C. Tu, D. Shin and D. Shin, ”A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms,” 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Chengdu, China, 2009, pp. 183-187, doi: 10.1109/DASC.2009.40.