Predicting H1N1 and seasonal flu vaccine uptake from sociodemographic characteristics and perceived vaccine effectiveness: a machine learning-based approach

Bethany Denton
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxbd2@nottingham.ac.uk

Abstract— **Vaccine hesitancy (a reluctance or refusal of vaccination despite the availability of vaccination services) has been a longstanding barrier that public health campaigns have struggled to overcome. Recent years have seen the re-emergence of viruses in countries where they were thought to have been eliminated, such as measles reappearing within the United Kingdom and USA. This, combined with the world-altering impact of the COVID-19 pandemic, has shown the consequence that decreased vaccine uptake can have upon public health at a national and global level. It has highlighted the need for accurate predictive models, which can be used to inform public health officials and allocate health care resources efficiently.**

Predictive models are most useful in a real-world application when they can provide accurate predictions based on data that are easy and reliable to collect. However, self-reported behavioral data are more likely to contain inaccurate values, as they require the participant to honestly assess their own behavioral patterns. This paper therefore aims to determine whether data from the National 2009 H1N1 Flu Survey can be used to train a model for predicting vaccine uptake, without the use of the behavioral features collected in the original survey.

Multiple modelling approaches are compared; best results were seen using an ExtraTrees model, and with casewise deletion being used to deal with missing values. This model predicted whether participants had received either the H1N1 or seasonal flu vaccination with an AUC ROC score of 0.85, and an F_1-score of 0.79.

This paper shows that collecting self-reported behavioral data may not be necessary to create well-performing predictive models for vaccine uptake. This information can be used to better inform the design of future surveys investigating vaccination patterns. We also provide an adequate model for predicting H1N1 and seasonal flu vaccination uptake within a population.

Keywords—National 2009 H1N1 Flu Survey, H1N1 and seasonal flu vaccine uptake classification, decision tree and ensemble machine-learning model

I. Introduction

The World Health Organisation declared a pandemic in June 2009, caused by the H1N1 influenza virus (commonly referred to as ‘swine flu’ by media outlets). Shortly after the H1N1 flu vaccine became available in October 2009, the United States conducted the National 2009 H1N1 Flu Survey. The collected data contained self-reported information about respondents’ sociodemographic background, their opinions about vaccine effectiveness, as well as the behaviours they take to avoid disease transmission. Data was also collected on whether the respondents had received H1N1 or seasonal flu vaccinations.

A high proportion of immunisation in a community, provided through vaccination, will cause a decline in the spread of that disease. This is known as ‘herd immunity’, and as the contagiousness of a disease increases, the proportion of immunised individuals required to stop the spread of disease (the herd immunity threshold) also rises. Improving vaccine uptake is therefore an important strategy for minimising the spread of an infection during a pandemic, protecting the health of the population, especially those who are unable to be vaccinated.

Classification models that can predict vaccination uptake are a useful tool for guiding public health officials, helping to identify populations that may be more hesitant to be vaccinated and therefore more at risk of increased viral propagation. Many modelling approaches can also be used to recognise the features most important in predicting an individual’s likelihood of vaccination, providing further valuable insight to public health experts. If a person’s likelihood to receive a vaccine is related to their sociodemographic characteristics, then this may identify communities that would be particularly vulnerable to high infection rates during a pandemic, which could advise the distribution of healthcare resources more appropriately. If vaccine uptake is related to a person’s opinion or knowledge about the infection and its vaccine, then that could help to prioritise the need for a public education campaign regarding these topics.

However, the predictive power of a model is limited by the quality of data it receives. The National 2009 H1N1 Flu Survey dataset contains several attributes based on the participant's self-reported behaviours regarding hygiene and disease avoidance. Self-assessment of behaviour has an increased risk of dishonesty or inaccuracy, and the context of the questions leaves them more susceptible to potential bias if the survey interviewers fail to create a neutral and nonjudgemental environment. Data regarding behavioural patterns are also limited in their generalisability, due to how these behaviours can be influenced by cultural expectations, as well as differences in the local laws put in place during larger disease outbreaks. In contrast, sociodemographic data (e.g., age group, marital status) are easier to collect and hold less subjectivity.

Motivated by the limitations in behavioural data, this paper aims to create a classification system for vaccination uptake using only sociodemographic information and features related to the participants' opinions on vaccines. If a predictive model can be successfully created using this subset of features, then this information can be used to guide the design of future surveys collecting vaccination uptake data. It also increases the system's potential in real-world application, by creating a model that relies on data that is easier to obtain and more generalisable between countries.

II. Literature Review

The widespread societal impact of the Coronavirus pandemic has reiterated the importance of understanding what factors influence vaccine uptake. It has been shown that H1N1 and seasonal flu vaccine uptake correlates with later COVID-19 vaccine uptake [1], highlighting the potential for the National 2009 H1N1 Flu Survey dataset to provide useful insight into current vaccination strategies.

Several papers work to identify the factors that influence vaccine uptake; understanding these variables provides a first step in uncovering the barriers that prevent people from getting vaccinated, allowing public health officials to design more effective and targeted strategies. Brien et al. [2] provides a systematic review regarding A/H1N1 vaccine uptake and associated predictive variables, reviewing twenty-seven studies across twelve countries. They found several demographic factors (sex, age, and education level), as well as a positive opinion of the vaccine, to be influential in vaccine uptake. A similar review [3] supports these findings, as well as identifying race as a potential influencing factor. Additional literature not covered in these reviews provides further evidence of the potential importance of sociodemographic factors [4-8] and vaccine-related opinion [5-6, 8, 9, 10] in relation to vaccine uptake.

Many models have been created with the aims of better predicting vaccination uptake; here, we consider three such models, which all focus on vaccination for a different disease.

Cheong et al. [11] investigated vaccine uptake for COVID-19, using sociodemographic data compiled from multiple United States data sources. They created a supervised regression model using XGBoost, with a k-fold cross-validation accuracy score of 62%. Their model identified location, education, ethnicity, and income as significant features in their predictive model.

Hasan et al. [12] investigated vaccine uptake for measles, using health and demographic data from several Bangladesh data sources. Multiple machine learning models were tested (Naïve Bayes, random forest, decision tree, XGBoost, and LightGBM). They found an ensemble of LightGBM and XGBoost provided the best performance, yielding an accuracy of 80%. The team were able to create an accurate model using only 3-5 attributes.

Ayachit et al. [13] investigated vaccine uptake for H1N1 and seasonal flu, using health, behavioural and sociodemographic data from the United States' National 2009 H1N1 Flu Survey. They tested several machine learning models (MlBox, TPOT, random forest, MLP, linear regression, decision tree, polynomial feature, XGBoost, and CatBoost). They found the CatBoost model performed best, with an accuracy of 86%.

III. Methodology

A. Data Source

This study was conducted utilizing data from the National 2009 H1N1 Flu Survey, provided by the United States National Center for Health Statistics, and made available by DrivenData for their Flu Shot Learning competition [14]. Data was collected through a phone survey; responses from 26,707 adult participants are analysed in this paper. Of the 35 features collected, we evaluate the 25 that focus on sociodemographic factors as well as participant opinion on the illnesses' risk and vaccines' effectiveness.

B. Exploratory Data Analysis

The test set was isolated used an 80-20 training-test split of the dataset, using stratified sampling based on H1N1 flu vaccination status to account for the class label imbalance. The following steps only involved the training dataset, until the final model validation (Fig. 1).

Duplicate entries were removed. No outliers or data entry errors could be found during data exploration. The dataset was explored through multiple avenues - including correlation analysis and data visualisation - allowing us to see the correlation relationship between different variables, as well as the differing distributions for data features across the target labels, and gain insight into how the features relate to one another. Attribute transformation was also iteratively experimented with.

C. Data Preparation

The survey dataset contained a substantial number of unexplained missing values, which need to be removed or imputed to develop a robust classification model. Two strategies were tested for dealing with missing data. The first approach was case-wise deletion to drop rows that contained missing values. The second approach was imputation of the most frequent value (as all attributes analysed were categorical features). Attributes were encoded using OrdinalEncoder for ordinal features, and OneHotEncoder for nominal features.

D. Data Modelling

The problem is an offline, multilabel classification task. Due to the size of the dataset and the lack of continuous data input, batch learning is suitable. Several quick models were iteratively

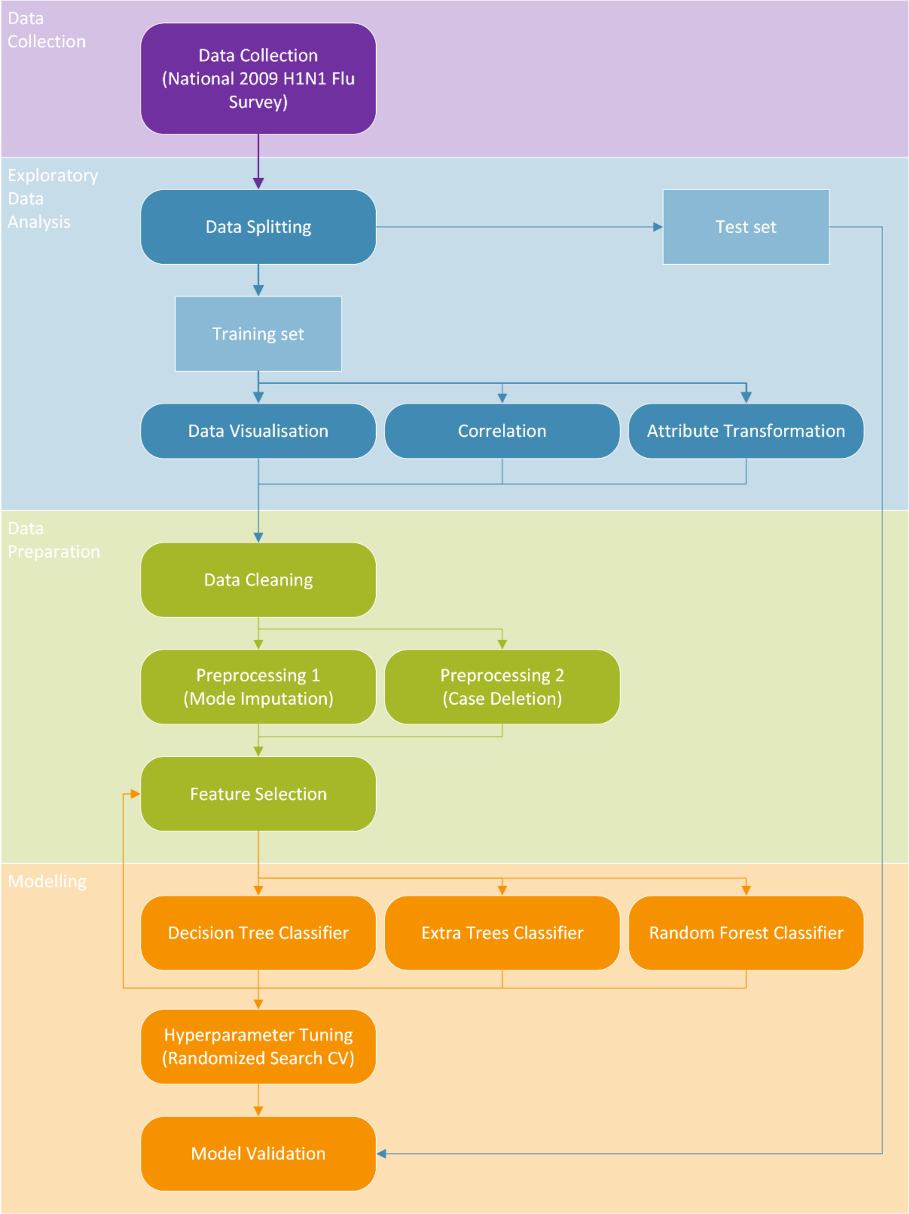

Fig. 1. Diagram of methodology used, including the data analysis, pre-processing and modelling steps taken

trained and compared, with features being dropped if they provided no useful information to the model. These models were created using the Python sklearn package.

Decision Tree Classifier: This model was selected for testing due to its strengths as a white box model. The model deals with sensitive data – such as sociodemographic factors – so if the model was ever used for guiding public health policies, then it would be advantageous if the model is explainable and if decisions based upon its output are easy to justify.

Random Forest / Extremely Randomized Trees (Extra Trees) Classifier: A major drawback in using a decision tree classifier is the model's high variance, and the increased potential for overfitting. An ensemble approach would help to overcome this, so both Random Forest and Extra Trees Classifiers were tested. However, the problem with using an ensemble approach is that most of the interpretability of the Decision Trees is lost in the process.

Due to processing limitations, RandomizedSearchCV was the chosen approach for hyperparameter tuning. The performance of the model was then tested using the best performing hyperparameters and the unseen test data.

Due to the skew in the labels, confusion matrices were used to visualize the performance of each model and where incorrect predictions were being made. The precision, recall and F1 score of each model were evaluated. Precision was valued higher than recall, as it is more important that the model does not overpredict the vaccination rate in a population. This is because if the model was used to guide public health policies, overprediction of vaccination rate could result in fewer resources than needed being distributed to vulnerable populations. AUC ROC scores were also calculated for the purpose of comparing scores to previous literature.

IV. Results

This section is broken into sub-sections to showcase the results obtained in each step of the methodology. Section IV.A will cover the insights gained during exploratory data analysis. Section IV.B will study the effect of different processing methods for handling missing values, as well as hyperparameter tuning. Section IV.C concludes by examining the performance of the tested models on the isolated test dataset, evaluated by reviewing their precision, F1 scores, and AUC ROC scores.

A. Exploratory Data Analysis

Depending on its tuning, a model such as Random Forest has the potential to become computationally expensive. One method to lower this computational cost is to remove features that are not adding substantial value. For example, when features are highly correlated with each other, it becomes possible to drop a subset of them, as the information they confer can be gathered from the remaining variables that they are correlated with. For example, it was shown that health insurance was weakly correlated with home ownership, poverty status, and unemployment (Fig. 2). Although this only captures possible linear relationships between the features, it is a good starting point for future exploration.

It is also useful to explore the relationship between the features and the labels, to gather some insight into which features have the highest potential for helping the model to predict vaccination status. Visualizing these correlations allowed us to quickly see the potential predictive value for features such as health insurance, perceived vaccine effectiveness and perceived illness risk. For nominal variables, the difference in distribution was visualized between positive and negative vaccination cases for each of the binary labels. This allowed similar insights to be gained; it can be seen how participants over 65 years old are more likely to have received a vaccination than those under the age of 35 (Fig. 2).

Correlation can also be seen between the dataset's labels, suggesting that the labels are not fully independent of each other, which is important to know for future methodology.

B. Pre-processing and preliminary modelling

15 features were used in the final models. The sociodemographic features used were age group, employment occupation, health insurance, household annual income with respect to 2008 Census poverty thresholds, and number of other adults and children in the household. The opinion-related features used were level of concern about H1N1 flu, level of knowledge about H1N1 flu, opinion of H1N1 and seasonal flu vaccine effectiveness, concern about getting sick from taking the H1N1 or seasonal flu vaccine, concern about getting sick from either H1N1 or seasonal flu if not vaccinated, and an overall 'H1N1 flu opinion' score calculated from their other H1N1 flu-related responses.

There were four target labels in the final system; whether participants had received the H1N1 flu vaccine, whether they had received the seasonal flu vaccine, whether they had received either vaccine, or whether they had received both vaccines.

The first approach for dealing with missing values was casewise deletion. The resulting training dataset contained only 9,934 entries. It is noted that this approach may be limited, as the missing values do not appear to be missing completely at random; if certain demographics or responses are more likely to be missing, then the model generated from the new dataset may not perform well when those values are present.

The second approach was imputation of the most frequent value (as all attributes analysed were categorical features). Although this appears less destructive, in those features that contain a high proportion of missing values, the most frequent value may become over-represented.

Optimal hyperparameters were calculated for each model being tested (Table 1). The method of processing missing values was also treated as a hyperparameter.

TABLE I. Hyperparameter values used with test dataset during model validation. Data columns represent the different models evaluated: Decision Tree (DT), Random Forest (RF), and Extra Trees (EXT).

Hyperparameter	**Casewise Deletion**			**Mode Imputation**		
	DT	***RF***	***EXT***	***DT***	***RF***	***EXT***
Maximum Depth	9	14	14	13	13	14
Maximum Features	13	14	14	13	14	14
Minimum Samples at Leaf	3	3	3	8	5	3
Minimum Samples for Split	5	9	5	9	2	4
Number of Trees	-	207	267	-	252	249

C. Classification

Performance measures allow us to evaluate each of the models. Model accuracy was between 75%-81% for the H1N1 vaccine label, 71%-78% for the seasonal flu vaccine label, and 70%-78% for the any vaccine label; however, due to the skew in some labels, this was not used as a performance metric. Instead, we looked at precision, F1 score (Fig. 3) and ROC AUC score (Fig. 4).

The casewise deletion approach is shown to perform better than mode imputation across all models. This may be due to the inclusion of the health insurance feature; although this feature provided a high predictive value to the model in exploratory analysis, it is imbalanced and has a high number of missing values, which would disproportionately affect simple imputation methods.

All models had higher precision and recall for predicting whether someone had received their seasonal vaccine compared to predicting whether someone had received their H1N1 flu vaccine. This may be due to the H1N1 vaccine class being substantially more imbalanced than the seasonal flu vaccine class.

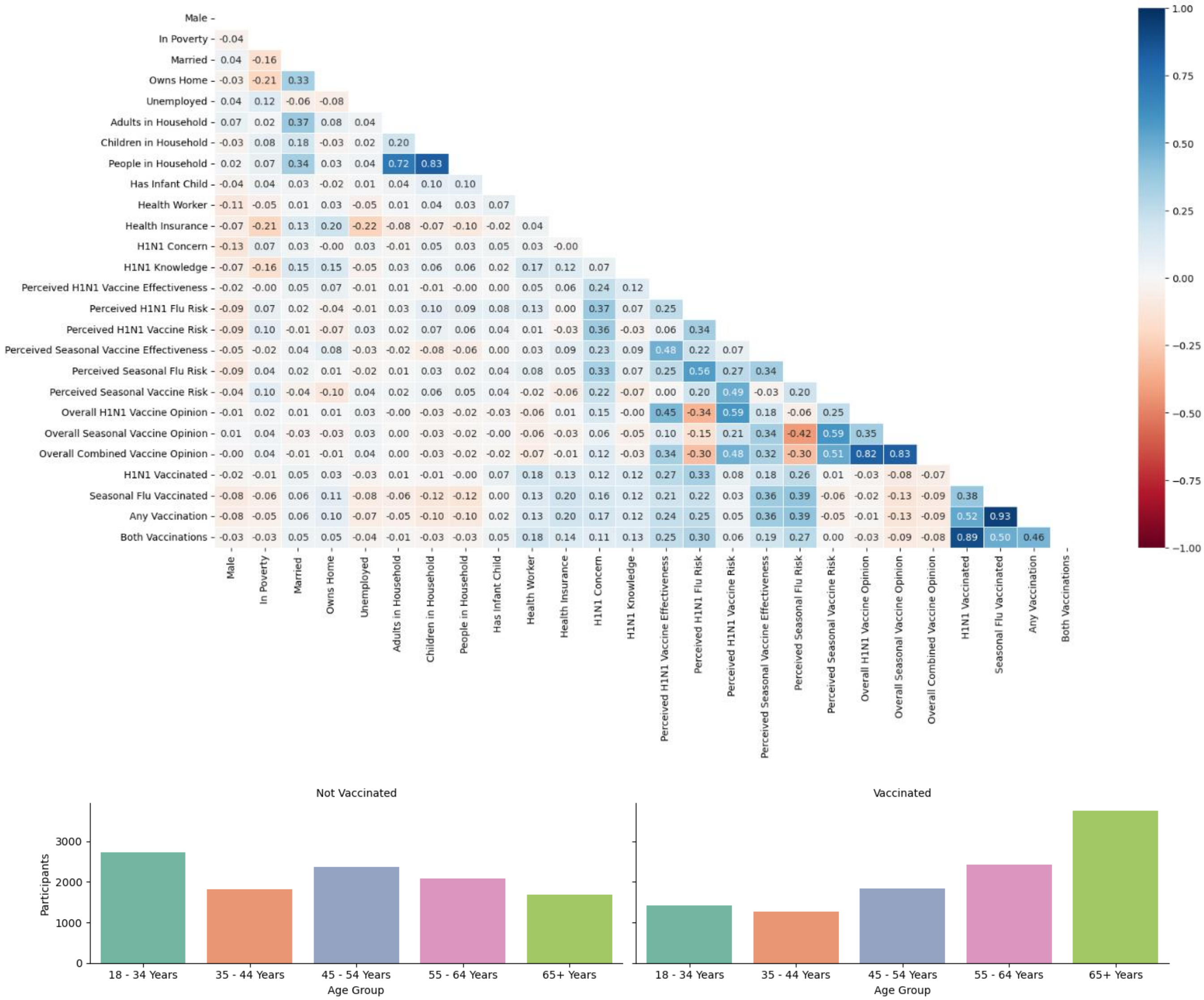

Fig. 2. A correlation matrix of all binary / numeric features and labels for the training dataset (top), and an example count plot, showing how the distribution for age differs between vaccinated and unvaccinated participants. Similar count plots were created for all categorical features (bottom).

It can be seen from the precision and recall scores for the H1N1 vaccine that the prediction mistakes being made are mostly false negative predictions. This was preferential to false positive predictions because, as explained above, false positive predictions have the potential to misguide public health policy in a way that prevents vulnerable populations getting the necessary resources. False negatives may also include people who had not received their H1N1 vaccination at the time of survey (as the vaccine had only become publicly available a few months prior) but went to receive it soon afterwards; these people would have a negative H1N1 vaccination label while having features associated with a positive vaccination label.

Both ensemble models (Random Forest and Extra Trees) performed better than the Decision Tree model in almost all metrics across all labels. The Random Forest model and Extra Trees model had similar performances; however, the computational runtime for Extra trees was noticeably faster.

V. Discussion

The separate pre-processing approaches were seen to have a large effect on the performance of the final model, with casewise deletion generally performing better across most cases. However, neither approach is without its limitations. As previously discussed, casewise deletion may cause longer-term issues with the model's performance, if the missing values in the training dataset were not missing completely at random. For example, if a certain demographic refused to give information regarding their health insurance, then that population would be dropped from the training dataset. The resulting trained model may then perform badly when trying to make predictions about that demographic in future datasets. On the other hand, mode imputation skews the distribution of features towards their most common value, which can cause problems for features with large numbers of missing values. Alternative approaches that could be tried in future studies include testing more advanced

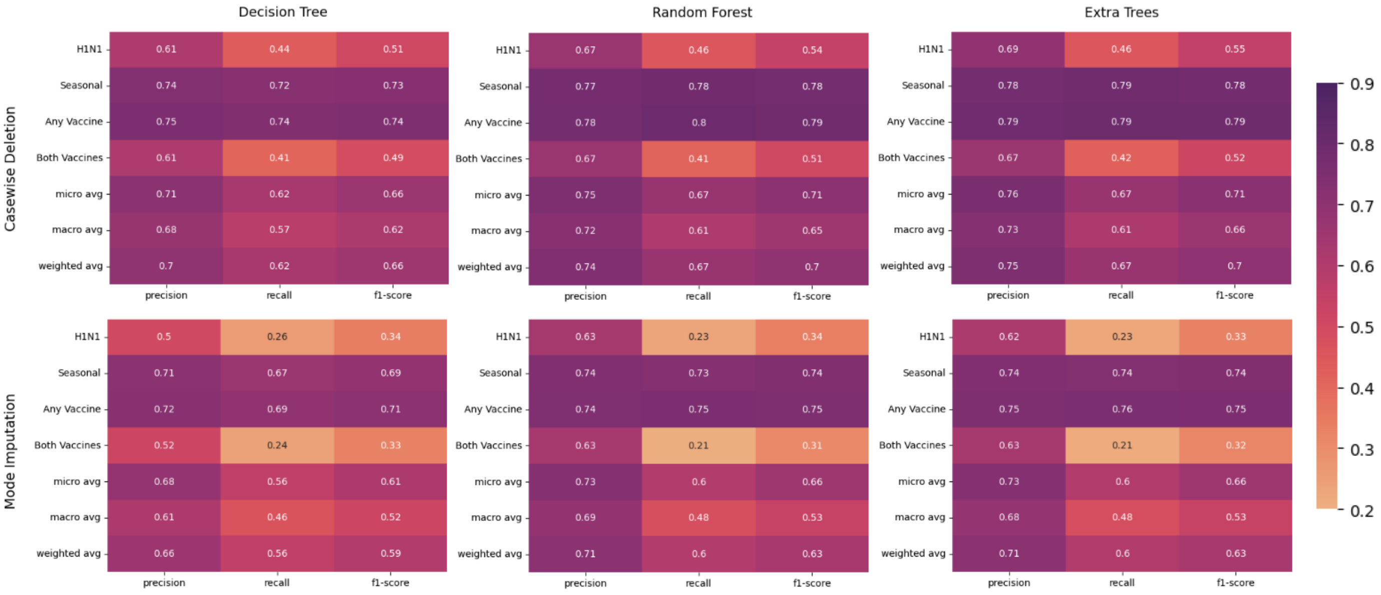

Fig. 3. Heatmaps to visualise the precision, recall and F1 scores of models across different pre-processing approaches and model types. Rows shows values for each label, as well as the micro average (micro avg), macro average (macro avg), and weighted average (weighted avg).

imputation methods, or only selecting features that have a maximum threshold for number of missing values.

The different classifier models used had a smaller effect on the performance of the model. The ensemble methods had higher precision scores and AUC ROC scores compared to the Decision Tree model. This is likely due to the tendency for Decision Tree models to have high variance and overfit the training data. In comparison, both Random Forest and Extra Trees models introduce extra randomness during tree design, only considering a random subset of features during node splitting. By doing this multiple times and comparing the generated decision trees, these models introduce a lower variance.

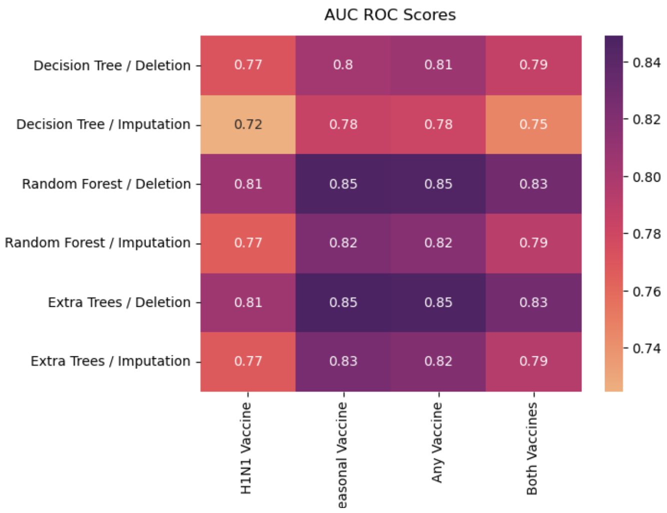

Fig. 4. Heatmap to visualise the AUC ROC scores for each combination of model and pre-processing approach. Columns show the value for each label.

Both the Extra Trees and Random Forest models perform similarly in performance metrics; as such, future users would be able to choose whether they prefer the faster computational speed of the Extra Trees model or the slightly increased performance of the Random Forest model. However, there is a large downside to these models. The high interpretability of a Decision Tree model is lost in the ensemble methods; as such, if the user values being able to explain the decisions made by the model (for example, if they are designing public health policies), then the Decision Tree model may prove optimal for them despite its slightly lower predictive performance.

The models outlined in this paper gave a similar level of performance to the models seen in previous literature.

Cheong et al. [11] reported an accuracy of 62% with their XGBoost model for predicting COVID-19 vaccination, which is lower than the accuracy achieved in our models. Some of the features identified as significant in their model (location, education, and race) did not provide much predictive value in our models and were not used as features in the final model. The only feature that proved valuable for both them and us was household income. Their datasets had more precise geographical location data compared to the 2009 H1N1 Flu Survey data; it would be interesting to explore whether it was this additional precision that allowed the feature to provide a high predictive value to their model.

Hasan et al. [12] reported a precision of 0.78 and an AUC score of 0.725 for their Random Forest model in measles vaccine prediction. Comparatively, the Random Forest model in this paper achieved a weighted average precision of 0.74 and an AUC ROC score between 0.81-0.85 across the various labels. However, Hasan et al. also designed a model using an XGBoost-based feature selection method and a LightGBM classifier, giving a precision of 84.60% and an AUC score of 80%. In addition, they were able to do this using only 3-5 features. In comparison, the model in this paper used 15 features in its performance. Differences seen may be due to the different model designs; however, they may also be due to the differences in dataset populations (Bangladesh vs the United States), and the differences between the diseases (measles vs flu varieties).

Ayachit et al. [13] worked on the same National 2009 H1N1 Flu Survey dataset, allowing for more direct comparisons between results. Their model aimed to predict whether participants had received either of the vaccines, giving consideration to all the available features. Their Random Forest model achieved an accuracy of 0.793 and their Decision Tree model achieved an accuracy of 0.80. This accuracy is higher than the accuracy achieved in this paper's models. Their best performing model used the CatBoost library to achieve an accuracy of 0.86. Other performance metrics were not included in the paper regarding the Random Forest model, making it difficult to assess both the full performance of their model and whether the inclusion of behavioural features improved the model in terms of more valuable performance metrics.

VI. Conclusion

This paper creates a classification system for predicting H1N1 flu and seasonal flu vaccination, without the need for behavioural data. Of the parameters covered in this paper, it is concluded that the best overall performance was seen with a casewise imputation approach and an Extra Trees Classifier model; however, if interpretability is important, a Decision Tree Classifier was shown to still perform well. These findings are limited by the impact that casewise deletion may have on generalising the model outside of the dataset. The performance of the model for predicting H1N1 flu vaccination status also has room for improvement; if it was possible to collect the data, it would be interesting to assess how many nonvaccinated participants received their H1N1 flu vaccination after the survey data were collected. It would also be of interest to see how the model performs when the behavioural features are included in the process, so that a direct comparison can be made between the models, and the importance of behavioural features can be better quantified.

References

[1] S. Brien, J. C. Kwong , D. L. Buckeridge (2012), "The determinants of 2009 pandemic A/H1N1 influenza vaccination: a systematic review," Vaccine, vol. 30, issue. 7, pp. 1255-1264, February 2012.

[2] P. Nair and D. P. Wales, "Seasonal and 2009 pandemic H1N1 vaccine acceptance as a predictor for COVID-19 vaccine acceptance, " in Cureus, vol. 14, issue. 1, e21746, 2022.

[3] A. Bish, L. Yardley, A. Nicoll, S. Michie, "Factors associated with uptake of vaccination against pandemic influenza: a systematic review," in Vaccine, vol. 29, 38, pp. 6472-6484, 2011.

[4] A. E. Burger, E. N. Reither, S. Mamelund, S. Lim, "Black-white disparities in 2009 H1N1 vaccination among adults in the United States: a cautionary tale for the COVID-19 pandemic," in Vaccine, vol. 39, issue. 6, pp. 943-951, 2021.

[5] J. Nikolovski, M. Koldijk, G. J. Weverling, J. Spertus, M. Turakhia, L. Saxon, M. Gibson, J. Whang, T. Sarich, R. Zambon, N. Ezeanochie, J. Turgiss, R. Jones, J. Stoddard, P. Burton, A. M. Navar, "Factors indicating intention to vaccinate with a COVID-19 vaccine among older U.S. adults," in PLoS ONE, vol. 16, issue. 5, e0251963, 2021.

[6] Y. K. J. Han, S. Michie, H. W. W., Potts, G. J. Rubin, "Predictors of influenza vaccine uptake during the 2009/10 influenza A H1N1v ('swine flu') pandemic: results from five national surveys in the United Kingdom," in Preventive Medicine, vol. 84, pp. 57-61, 2016.

[7] K. O. Kwok, K. Li, W. I. Wei, A. Tang, S. Y. S. Wong, S. S. Lee, "Influenza vaccine uptake, COVID-19 vaccination intention and vaccine hesitancy among nurses: a survey," in International Journal of Nursing Studies, vol. 114, e103854, 2021.

[8] E. M. Galarce, S. Minsky, K. Viswanath, "Socioeconomic status, demographics, beliefs and A(H1N1) vaccine uptake in the United States," in Vaccine, vol. 29, issue. 32, pp 5284-5289, 2011.

[9] M. de Vries, L. Claassen, M. Lambooij., K. Y. Leung, K. Boersma, A. Timen, "COVID-19 vaccination intent and belief that vaccination will end the pandemic," in Emerging Infectious Diseases, vol. 28, issue. 8, pp. 1642-1649, 2022.

[10] A. R. Ashbaugh, C. F. Herbert, E. Saimon, N. Azoulay, L. Olivera-Figueroa, A. Brunet, "The decision to vaccinate or not during the H1N1 pandemic: selecting the lesser of two evils?" in PLoS ONE, vol. 8, issue. 3, e58852, 2013.

[11] Q. Cheong, M. Au-yeung, S. Quon, K. Concepcion, J. D. Kong, "Predictive modeling of vaccination uptake in US counties: a machine learning–based approach," in Journal of Medical Internet Research, vol. 23, issue. 11, e33231, 2021.

[12] K. Hasan, T. Jawad, A. Dutta, A. Awal, A. Islam, M. Masud, J. F. Al-Amri, "Associating measles vaccine uptake classification and its underlying factors using an ensemble of machine learning models," in IEEE Access, vol. 9, pp. 119613-119628, 2021.

[13] S. S. Ayachit, T. Kumar, S. Deshpande, N. Sharma, K. Chaurasia, M. Dixit, "Predicting H1N1 and seasonal flu: vaccine cases using ensemble learning approach," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 172-176.

[14] DrivenData, "Flu shot learning: predict H1N1 and seasonal flu vaccines," 2023, [online] Available: https://www.drivendata.org/competitions/66/flu-shot-learning.