Behavioural Insights into H1N1 and Seasonal Vaccine Refusal

Daniel Pereira De Abreu
School of Computer Science
The University of Nottingham
Nottingham, United Kingdom
psydp1@nottingham.ac.uk

Viktoria Stoeva
School of Computer Science
The University of Nottingham
Nottingham, United Kingdom
psyvs1@nottingham.ac.uk

***Abstract*— The primary purpose of this report is to answer two questions showing how the behavioral trends of individuals and groups influence their decision not to accept the H1N1 and Seasonal flu vaccines while elucidating and critiquing the development of classification models predicting vaccination rates. Since vaccines are an effective way to hinder the spread of infectious diseases, understanding the reasons behind people's refusal to get them could be utilized to drive their acceptance.**

The data used is from a phone survey collecting information on whether individuals have received the two vaccines, as well as related factors contributing to the decision. After applying data cleaning and wrangling techniques, a multitude of multi-label classification models were implemented. XGBoost emerges as the best performing model with a ROC score of 85% among various algorithms employing a pipeline that encompasses MinMaxScaling, feature selection and constant, median and mode imputation techniques.

Index Terms – Machine Learning, Vaccine Uptake, Data Science, Multi-Label Classification

I. INTRODUCTION

Individuals of all ages are at risk of developing severe illnesses from the H1N1 virus or seasonal flu. Vaccines prevent specific infectious diseases from spreading uncontrollably by enhancing your immune system. The H1N1 Influenza virus, more commonly known as Swine Flu caused a pandemic in 2009. A vaccine was developed that mitigated the spread and limited the overall impact of the disease. Seasonal flu vaccines safeguard individuals from the most common viruses for the upcoming flu season. Yet, there are various ways individuals justify refusing both the H1N1 and seasonal flu vaccines. "The 4 categories are religious reasons, personal beliefs or philosophical reasons, safety concerns, and a desire for more information from healthcare providers" [1].

A phone survey was conducted to collect the data on whether candidates had received vaccines for H1N1 or Seasonal Flu. This information is used as the dependent variables, which need to be predicted based upon the independent variables: demographic questions asked during the survey. The dataset has a total of 38 features with 2 of them being our target predictions. The remaining features are all categorical. There is a data imbalance with every instance having missing data and a skewed class for the H1N1 vaccine. Variables such as age_group, race and hhs_geo_region hold high imbalance which could lead to bias.

The aim of this report is to provide answers to two questions focusing on individual and group trends in order to increase mass immunity. What behavioural factors can be identified forecasting H1N1 and seasonal flu vaccine refusal among the different demographic groups, and how this insight can be used to increase vaccination rates? What are the underlying social and psychological factors influencing individuals who opt to receive the H1N1 and seasonal flu vaccines, and how can this be utilized to allow for better communication and therefore drive vaccine acceptance? Determining the probability of person receiving the H1N1 and Seasonal Flu vaccine based on specific factors is a multi-label classification task. It can be solved using various Machine Learning (ML) models to predict an outcome based on gathered data.

II. LITERARUTE REVIEW

Multiple research papers have been written using this particular dataset for machine learning model predictions. The first chronologically published of those is the 2020 "Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning approach" by Sai Sanjay Ayachit et al. In this paper their focus is on comparing 9 models in predicting the likelihood of an individual getting either or both vaccines in the upcoming season. They find that "CatBoost gave the best performance with an accuracy of 0.8617 followed by the XgBoost and MlBox" [2], both of which achieved around 85% accuracy.

In February 2021 the paper "FLU SHOT LEARNING: PREDICT H1N1 AND SEASONAL FLU VACCINES" was written by Srividya Inampudi et al. Their main objective was to also develop models that can successfully predict how likely it is for people to get vaccinated. Once again, multiple methods were used, with the highest achieved results being by an ANN with two hidden layers and an SGD optimizer. Yet, here they predict each vaccine separately, with the ANN attaining "accuracy over 82% in H1N1 flu vaccination prediction and 86% in Seasonal flu vaccination prediction" [3].

The published in November 2021 "H1N1 Vaccination Status Prediction" by Ruoyu Zhang, Tongyu Zhao and Yuhao Zhou paper focuses on identifying which "social, economic, and demographic characteristics are associated with personal vaccination patterns" [4]. They conclude that 5 factors have the highest importance when it comes to affecting the vaccination status of individuals: "seasonal flu vaccination status, opinion on H1N1 risk, doctor recommendation of H1N1, opinion on effectiveness of H1N1 vaccine, and doctor recommendation of seasonal flu vaccine" [4].

With researchers aiming to develop models that achieve new heights in prediction accuracy and therefore focusing only on feature subsets which aid that, a huge research gap is created. The relationship between the other features and the outcome should be analyzed further in order to discover hidden trends resulting in vaccine refusal. Should that show compelling evidence that people's social, psychological, and behavioral factors influence their vaccination acceptance, more efficient ways to communicate the benefits and address their fears could be created. This could in turn be applied to other medical disciplines and ultimately encourage trust in practitioners, resulting in potentially saving lives.

III. METHODODOLOGY

People who refuse to get vaccinated are not united under the same ideology or outlook. “One commonality, however, is that they have been subject to a confusing mix of sometimes contradictory information and misinformation about vaccines and vaccine safety” [5]. Considering this, while developing our classification models, we focalize on identifying the different factors that strongly influence vaccine refusal with the aim to address them in the future. Our project involves exploring and pre-processing the survey data, as well as implementing and comparing various multi-label classification models. Figure 1.1 below illustrates the entire development process.

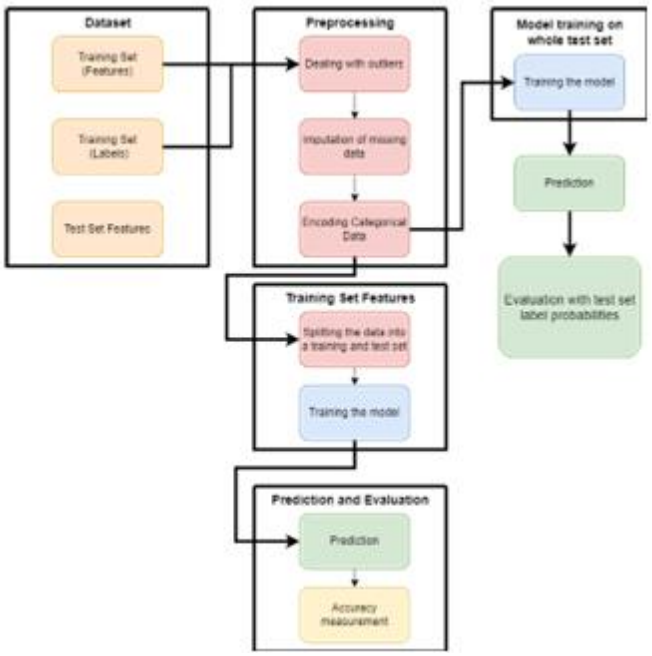

Figure 1 - Process Chart

We employ standard models such as Random Forest, Gaussian Naive Bayes, and K-Nearest Neighbours. In addition to those models, we incorporate two gradient boosting algorithms – the gradient boosted decision trees XGBoost and Catboost. To automate the data pre-processing section, pipelines from the SKLearn library are used for the majority of our models. Every model is created by defining a classifier chain that includes feature selection from mutual_info_classif, a scaler such as MinMaxScaler and the classifier itself. Another pipeline is then made for each model that combines its respective chain with the preprocessor pipeline. Every model, except Naïve Bayes has been implemented with custom parameters searched using RandomizedSearchCV.

A. Data Cleaning and Pre-Processing

At the initial stage, we load the two provided datasets containing features and labels and begin exploring the data. We output observations of it using functions such as head, tail and iloc to gain a better understanding of the data. We investigate the data further by applying the info and describe functions to the features. We observe that the number of instances varies for the different features with the range being from 14433 to 26688. Additionally, we see that the minimum value for all variables is 0 or 1 and the maximum value ranges between 1 and 5, with the 50th percentile being 1 for most variables.

To expand our understanding even further, we concatenate the two dataframes with concat and generate a profiling report using the ProfileReport from the pandas_profiling library. It provides us with a detailed summary for variables, including a countplot of each value per variable, with the option to explore more details which helps us identify outliers. As an alternative to, we also manually check if any duplicate and missing data is present, as well as outliers. The interquartile range (IQR) is used to find the outliers, showing that household_children and household_adults have such values. Yet, we decide against removing them as both appear to be misrepresented data. Additionally, no duplicates are found. However, the dataset has a significant number of missing values across different features that are identified using the Pandas library. Therefore, data cleaning is deemed a priority. The instances of missing data are visualized using a matrix plot from the missingno library to aid the extraction of valuable insights. Calculating the percentage of missing data in each column shows that the features “health_insurance”, “employment occupation” and “employment_industry” exhibit the highest amount with all having over 40%. To avoid misleading results those features are dropped.

The rest of the missing data is handled with various imputation strategies. To handle the numerical and categorical features separately, column transformations are performed. Numerical data is scaled using the MinMaxScaler SKLearn class to ensure that each column has a mean of zero and unit variance. Immediately after this, SimpleImputer is utilized to fill in any missing data instances with different strategies accordingly. We impute the ordinal object data using a constant method strategy and filling the missing values with -1 to avoid bias and converse data. On the other hand, Boolean data has been imputed with its mode while ordinal numeric data has been imputed using the median of each column. One-Hot Encoding is then applied to encode each categorical feature into separate binary variables, providing us with an efficient representation of the categorical data for the purpose of modelling.

B. Visualization

To gain insights and better understand the relationships and patterns in the dataset, various visualization techniques are employed. These methods include class balance and label correlation analysis, cross-tabulation and feature correlation, examination of feature distributions, visualization of outliers, amongst others. The following subsections will outline the observation of each of these techniques and their implications for the classification models.

1) Class Balance and Label Correlation

To check if the dataset is balanced, we examine the distribution of the two target variables with count plots. We use the countplot function from the Seaborn library and map one of the target variables from the labels dataset to the x axis of the graph. That provides us with a bar representing the number of observations in each category – accepted or refused the vaccine.

2) Feature Correlation

Multiple ways are used to analyze the correlation between the features. Initially we use the corrwith function to output the correlation between each vaccine and the rest of the features. Then to visualize it, we employ the correlation heatmap from the Seaborn library. We not only make it display the feature names and the correlation coefficients in each block, but also pick the color scheme viridis to make it easier to decipher due to the contrasting colors. Another feature correlation heatmap, yet this time it measures the nullity correlation. It is created using the Missingno Python library to explore the connections between the missing data. Its coefficients are in the range -1 to 1 where 0 means no relationship. While -1 specifies that when one of the values is present the other is missing 1 showcases the opposite. Finally, to gain a deeper understanding of the connection between the variables, a dendrogram is implemented to represent the whole data using the dendrogram function from the same library.

3) Behavioral Features Distribution

In order to compare the distribution of the behavioral actions of individuals, we create an array of those features using the Numpy library and sum each one. We visualize them with a pie plot through the Matplotlib package that takes in the number of times each action is said to have been done in the survey and the action names as labels for parameters.

4) Behavioral Features Distribution per vaccine outcome

To visualize the association between the vaccine rates and the different behavioral features. We've generated a set of subplots. With each one comparing the rates of both the H1N1 vaccine and seasonal vaccine based on a specific behavioral feature. We can observe which behaviors have a significant impact on the vaccination rates and provide insights into the factors that influence individuals' decisions regarding vaccination. This is implemented by extracting the behavioral features from a list and specifying rows and columns based on the length of the columns to plot. A loop then iterates through the list and plots are drawn on the corresponding subplots with a total of 14 plots.

5) Cross Tabulation

In order to examine the association between the behavioral features and the vaccines taken, we performed cross-tabulation. This was introduced by creating a crosstab function that normalizes the result by columns and outputs a table that summarizes the joint distribution of he variables specifies.

6) Gender Distribution

To present and analyze the gender distribution by vaccination status, we create stacked bar charts for each subplot where the data is organized in tabular form using an unstack() function, allowing it to be organized for further analysis. In order to make it suitable for visualization using seaborn, we "melt" the data, which transforms it into a shape that is suitable for the visualization to take place. Due to the large size of the graph, we use the tigh_layout() function to ensure the spacing is adjusted and the subplots are properly arranged. These side-by-side subplots allow a direct comparison of the gender distribution, aiding in supporting our findings with a concise summary of the data.

As an alternative, we group the dataframe by vaccine and sex and calculate the size of each group with the groupby and size functions. Then we use the unstack function to reshape the data for the plot into a tabular format which is more appropriate for the graph. And plot a stacked bar chart by picking the kind of the plot to be a bar, setting the stacked parameter to True. We also change the colours of the stacked sections to pink and steelblue and add a title.

7) Outlier Identification

To identify outliers, we created a function that is based on the interquartile range (IQR) method. The IQR method is a robust statistical approach which is less sensitive to extreme values in comparison to other methods such as the Z-Score method. The IQR function begins by calculating the first (Q1) and third quartile(Q3) of the input features, which equated to the 25^{th} percentile and the 75^{th} percentile of the data distribution. We then calculate the IQR by subtracting Q1 from Q3 which represents the range between the two, leading to it covering the middle 50% of our data. Based on this range, we set a condition as seen in Equation 1 which identifies the outliers and returns them as a series

$$\begin{gathered} below \\ Q1 - 1.5 \times IQR \\ or\ above \\ Q3 + 1.5 \times IQR \end{gathered}$$

Equation 1 - IQR condition

One of the ways implemented to visualize the outliers is using the boxplot function from Seaborn and passing it one of the variables with outliers. The other way of visualizing them is by using count plots with the same library. They are created through passing a feature as the x axis of the count plot function, which in return provides the number of instances for each value in that variable represented as different height bars.

C. Model Development

1) Gradient Boosted Decision Tree

XGBoost is a gradient boosting decision tree algorithm. It takes advantage of the collective knowledge of an ensemble of decision trees to predict the target variable. To optimize the performance of the XGBoost model and mitigate fact that it is prone to overfitting we tune its hyperparameters. The key hyperparameters controlling the model's behavior include the maximum depth of each tree, the learning rate, the number of estimators (the number of decision trees in the ensemble), the regularization parameter (reg_lambda), and the percentile for feature selection. In order to identify their optimal combinations, a randomized search approach is employed. This involves specifying a parameter grid containing a range of values for each hyperparameter. The RandomizedSearchCV function is then utilized, which exhaustively sampled combinations from the parameter grid and evaluated their performance using cross-validation, aiming to identify the set of hyperparameters that yields the best performance. The XGBClassifier library is used in the implementation of the XGBoost model and the fine-tuning is

performed by the RandomizedSearchCV from SKLearn, ultimately enhancing the overall result.

2) Random Forest

Random Forest is built upon multiple decision trees and combines their predictions to output a final prediction. This is implemented using the SKLearn RandomForestClassifier library to predict the target labels. The hyperparameters of the model, such as the number of trees, the maximum depth of trees, and the number of features to consider at each split, are tuned using SKLearn's RandomizedSearchCV to improve the model's performance. It works by randomly selecting hyperparameters from a defined search space and fitting a model to the training data using those hyperparameters. The model's accuracy is then evaluated on a validation set, and the process is repeated a specified number of times and provides the hyperparameters that lead to the highest accuracy result.

3) Naïve Bayes

Naïve Bayes uses Bayes' Theorem by assuming that all features are conditionally independent from each other. We've used the GaussianNB classifier provided by the SKLearn library. GaussianNB assumes that the likelihood of each feature given that the labels are Gaussian functions.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Equation 2- Gaussian Naive Bayes Formula

4) Multilabel K-Nearest Neighbour

The K-Nearest Neighbors algorithm is a versatile approach that can be applied to both regression and classification tasks. The algorithm identifies the k nearest neighbors to the given observation and assigns it to the class that appears most frequently among these neighbors. This is implemented using the SKLearn library as it provides the KNeighborsClassifier, which we use in combination with GridSearchCV to optimize the value of k that yields the best results. By tuning k, the model can be adjusted to suit the characteristics of the data and enhance its performance. The RandomizedSearchCV function allows for a systematic exploration of a range of k values to identify the optimal value that maximizes the classifier's accuracy.

5) CatBoost

The CatBoost algorithm is our last model, and the second gradient boosted decision tree, working similarly to XGBoost. It is implemented using the CatBoostClassifier which is compatible with SKLearn. CatBoost inherently encodes categorical features without us having to do it ourselves. The hyperparameters of the CatBoost model, such as the learning rate, max bins, growing policy, max depth and more, were tuned using RandomizedSearchCV to provide higher performance. This involvs randomly sampling the hyperparameters from a predefined search space and fitting the model to the training data, while evaluating its accuracy on a validation set. The process is repeated multiple times to find the hyperparameters that yield the highest accuracy result.

IV. RESULTS

A. Visualization Results

The implemented visualizations showed hidden patterns in the data. The distribution of the two target variables is examined. As depicted in Figure 2, it reveals that nearly half of the individuals have received the seasonal flu vaccine, while only 20% have been vaccinated for H1N1. In terms of class balance, it can be inferred that the distribution of individuals vaccinated for the seasonal flu is balanced, while the distribution of those vaccinated for H1N1 is imbalanced.

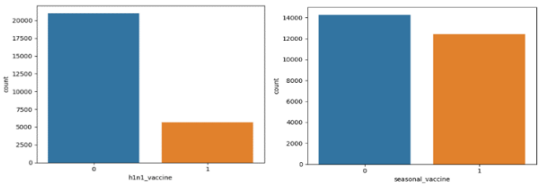

Figure 2 - Proportion of both H1N1 and Seasonal

From the feature correlation matrix in Figure 3, we observe that certain attributes exhibit moderate positive correlation, with two of particular interest being doctor_recc_h1n1 vaccine and doctor_recc_seasonal vaccine, which display a positive correlation of 60%. This finding implies that there is a strong relation between these two attributes, and they are highly relevant in predicting vaccine acceptance. Similarly, opinion_h1n1_risk, opinion_seas_vacc_effective and opinion_seas_risk seem to have a noticeable influence on the outcome. A distinct finding is that should a person get one of the vaccines, they are more likely to get the other one too.

To assess the behavioral insights, it can be observed that several of the behavioral-based features show mild positive correlations with each other. However, it should be noted that there is minimal correlation between most of the behavioral-based features and vaccine uptake, those being behavioral_antiviral_meds, behavioral_large_gatherings, behavioral_avoidance, behavioral_outside_home. Additional variables that don't make good predictors of vaccine acceptance based on the heatmap are child_under_6_months, household_adults, amongst others. The data also suggests that individuals who engage in behaviors such as not regularly washing their hands, not wearing face masks, and regularly touching their face are less likely to receive either the H1N1 or seasonal flu vaccine.

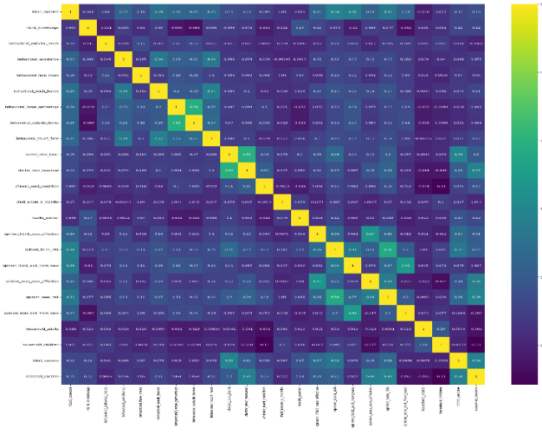

Figure 3 - Feature Correlation Heatmap

From the nullity correlation heatmap presented in Figure 4, we can conclude that there are no instances of variables whose absence of provided answer indicates that the value for another variable will be missing. Yet, some do have a strong positive correlation of both values appearing together, with the highest two correlations being household_adults with household_children and doc_recc_seasonal and doc_recc_h1n1. It is also worth noting that there is a strong relationship between a person's opinion on the risks of the disease and the effectiveness of the vaccine. And the behavioral features missing all have a 0 or near 0 correlation with any other values missing.

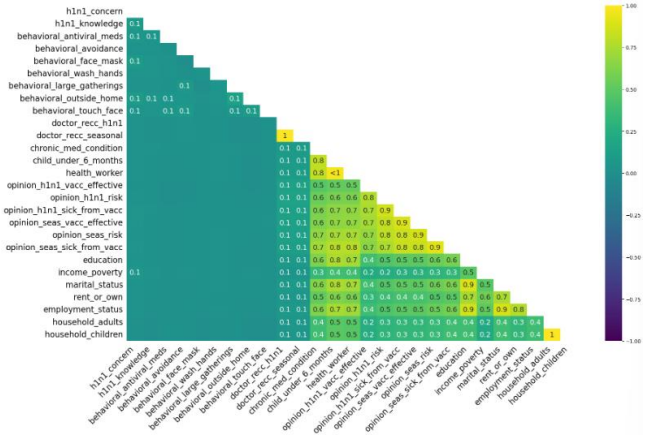

Figure 4 - Nullity Correlation Heatmap

The Dendrogram shown in Figure 5 represents the hierarchical clustering of the features and since here a top-down approach is used to make conclusions, we can see that several distinct groups are formed based on similarity. The behaviors are related with each other, similarly people's opinions on the disease and vaccines are related. Their personal choices about family, education and employment form a different cluster. It also shows that there is no clear clustering pattern between race, age_group, sex, region and census.

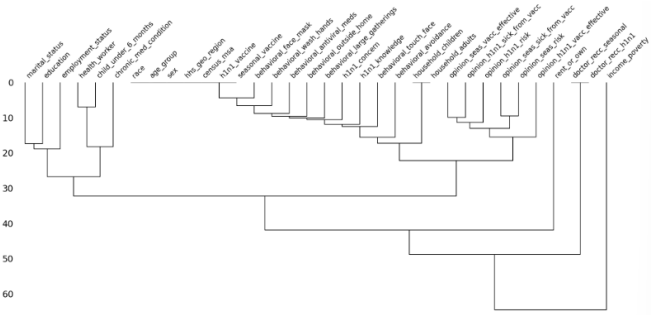

Figure 5 - Dendrogram

When investigating the behavioral features further, we observe the two least performed behavioral actions are wearing a face mask and taking antiviral medications. Not only that but when exploring their distribution per vaccine, it is evident that the h1n1 vaccine is refused more often than the seasonal one. Also, an individual performing behavioral actions increases the chance of them getting both vaccines, with the influence being greater for the seasonal one.

The same conclusion is reached using cross-tabulation tables, as the percentages of individuals taking the vaccine tends to increase with the number of positive behaviors. This confirms that people are more likely to accept vaccines when adopting the behavioral actions.

Upon analysis, it can be observed in Figure 6 that the H1N1 vaccine is refused more frequently than the seasonal one. Another conclusion reached is that the majority of the individuals who received the seasonal vaccine are female. Similarly, this trend is also observed with the H1N1 vaccine, suggesting that women may be more wary of infection than men. These findings highlight the potential importance of gender in relation to vaccination efforts and further investigation may be warranted to better understand the underlying factors contributing to this observation.

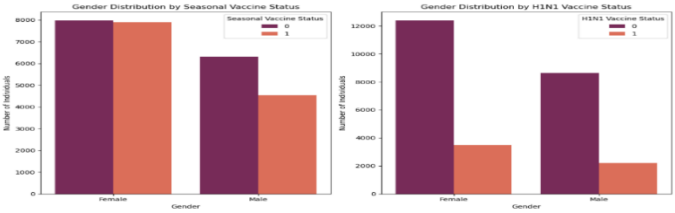

Figure 6 – Gender Distribution for Vaccination Outcome

B. *Model Results*

To evaluate the performance of the classifiers, we employ a variety of metrics that are standard for classification tasks. Specifically, we utilize the following techniques: confusion matrix, accuracy score and area under the receiver operating characteristic curve (ROC-AUC). The confusion matrix allows us to observe the distribution of correctly and incorrectly classified instances, assisting us in concluding if the model is predicting too many false positives or false negatives. False positives occur when the model predicts a positive outcome incorrectly, while false negatives happen when the model predicts a negative outcome incorrectly. Additionally, the accuracy score allows us to measure the proportion of instances that are correctly classified against the test sets, whereas the ROC-AUC metric technique measures the ability of the models to distinguish between the two classes by plotting the true positive rate against the false positive rate at various thresholds.

Model	ROC AUC Score
XGBoost	85.0%
Random Forest	84.3%
K-Nearest Neighbors (K=150)	83.3%
CatBoost	82.0%
Gaussian Naïve Bayes	63.6%

Table 1 - Models ROC AUC Scores

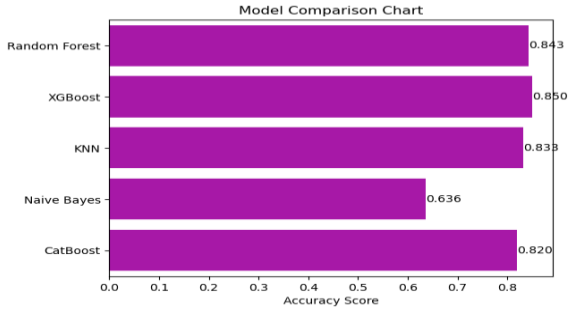

Figure 7 - Models Comparison Chart

Based on these results in Figure 7, it is clear that XGBoost performed exceptionally well in comparison to the other models tested. With a score of 85.0% being our highest and the lowest being 65.6% from using Gaussian Naïve Bayes. Additionally, the ROC curve in Figure 8, illustrates the results achieved by the XGBoost Classifier for both the H1N1 and seasonal vaccine labels, with AUC scores of 0.8459 and 0.8538, respectively. The ROC curve is closer to the top-left corner of the plot than our other models, and therefore further from the diagonal line which is also known as the "random guess line". That indicates that the XGBoost model is the most effective at making accurate predictions.

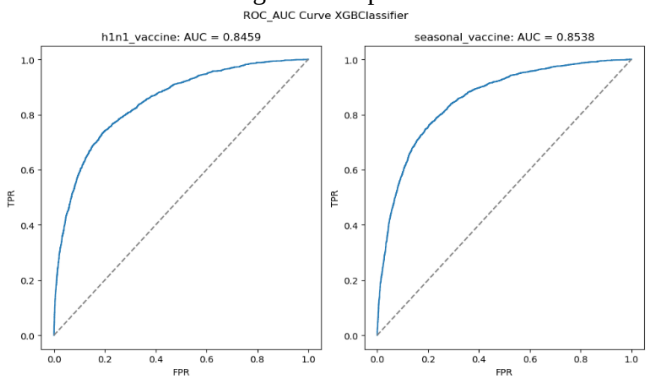

Figure 8 - ROC AUC Curve for XGBoost

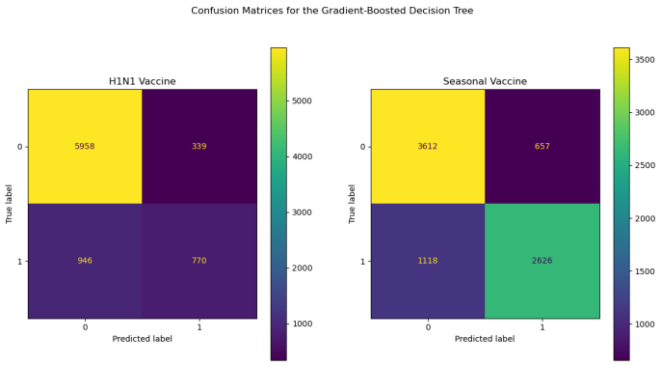

Figure 9 4 - Confusion Matrix for XGBoost

The XGBoost confusion matrix in Figure 9 shows that the model achieves the lowest number of false positives (339) and false negatives (946). By minimizing both types of errors, the XGBoost model demonstrated a good balance between precision and recall, resulting in accurate predictions and reliable performance. This indicates the model's effectiveness in distinguishing between labels and making accurate predictions.

V. Discussion

A. Comparison of Development Alternatives

By using the head and tail functions to output the first and last few rows of the dataset, we get insights into the overall structure and general trends of the data we would be using. On the other hand, the iloc function provides us with a detailed examination of the central records when specified with a segment of the dataset. By employing both exploration methods together, we get a comprehensive understanding of the dataset enabling us to make better decisions in the data preprocessing and modelling steps.

We attempted three methods for combining the features and labels dataframes: merge, append and concat. We found that while merge offers flexible join options, it presented challenges for us with duplicate values appearing and alignment issues. Append allows vertical stacking, however, it struggles with columns not present in both data frames which made it unusable. Ultimately, we ended up choosing concat, as it seamlessly concatenated the data frames horizontally, while preserving column alignment and ensuring no duplication is created.

When exploring the missing data, sum proved to be more beneficial to our research than any, as the function any returns a Boolean, while sum returns the number of missing values per column. When visualizing the missing data, the Missingno matrix represents each feature as a vertical bar and each missing value as a blank space within it. While the bar visualization from the same library provides a count of the missing values above each variable bar, allowing for more precise decision making.

Features with missing data exceeding 40% are dropped instead of being imputed, to avoid the introduction of bias and distortion of the distribution of the data. Although that decision could lead to loss of information, imputation of such variables would lead to unreliable prediction results.

We tested different imputation methods and settled on using ones based on the type of missing data present. For ordinal object data, a constant fill value of -1 is used to preserve the predefined order without disrupting the ordinality of it. Mode imputation is applied to the Boolean data to maintain the dominant pattern of the most frequent value, whereas median imputation is chosen for ordinal numeric data to account for the skewed data and preserve the overall distribution of the data.

When conducting outlier identification, some were found. However, we decided against removing them from the dataset, as although they are falling outside the expected range, they may not necessarily be incorrectly recorded. We believe they represent unique cases that should be retained to capture the full range of the data distribution and avoid potential loss of important data.

Two options for transformations were explored – Standardization and Normalization. We concluded that Standardization would be more beneficial in this case as we decided not to remove the outliers and it is less sensitive to them.

One-hot encoding is chosen over other encoding techniques such as label encoding, ordinal encoding, and binary encoding, due to its ability to represent categorical features without introducing an ordinal relationship. Label and ordinal encoding can potentially misrepresent the data by creating a wrong hierarchy which would lead to bias. The binary encoding is useful for reducing dimensionality; however, it assumes that categories have a specific hierarchy which may not be true.

For feature selection, SelectPercentile with mutual information is chosen over SelectKBest. SelectPercentile allows us to select features based on measuring the dependency between features and captures any non-linear relations and interactions. This ensures that there is a broader range of meaningful features and avoids the limitation caused by SelectKBest of having a fixed feature selection. We also choose it over picking the feature subset through the heatmap correlations as it minimizes redundancy.

B. Model Results Explanation

We observe notable differences in the performance of the XGBoost and Naïve Bayes models. XGBoost demonstrated the best performance, this was expected due to its ability to handle complex relationships and interactions while paired with the feature selection method chosen. Additionally, its ensembled structure that combines multiple weaker classifiers, leads to it effectively learning from the training data more efficiently and capturing both linear and nonlinear patterns. Naïve bayes is considered a much simpler model as it assumes that features are independent from one another, with no influence on each other. As expected, it is our worst performing model, as it struggles to capture the complex relationships present and needed to predict the vaccine labels. On top of this, XGBoost has optimal hyperparameters searched for, whereas, Naïve Bayes does not have such an option.

It is also worth noting that XGBoost aims to sequentially correct the errors made by previous trees, thereby improving the overall predictive performance. That makes it better than individual decision trees, which may suffer from limitations such as high bias or variance.

Random Forest also combines the predictions of decision trees. Due to the averaging effect of having multiple decision trees improved overall accuracy and also making it less prone to overfitting.

On the other hand, the performance of the KNN model approach is heavily influenced by the choice of k; a larger value of k can suppress the effects of noise in the data but may also make the decision boundary less clear and affect the ROC scores. Additionally, going over 150 neighbors does not improve the score further, it becomes stagnant.

The achieved results are similar to the 2020 findings of Sai Sanjay Ayachit et al [2], where the classification accuracy of their highest-performing model CatBoost was 0.86%, followed by XGBoost at 85%. Those are our leading models too, yet our XGBoost outperforms the rest. In addition to that, there is a similitude with the 2021 Srividya Inampudi et al observations [3] where their models are predicting the Seasonal flu vaccine slightly better than the H1N1 one too.

C. Research questions evaluation

Multiple conclusions were reached while exploring the behavioral, psychological, and social features in depth to answer the two research questions. Understanding the relationship between those and the vaccination outcome led us to discover multiple factors that influence vaccine refusal. Although, the people in the survey are not unified in any of them, hidden trends are identified.

One of those findings is that as positive behaviors decrease, the percentage of people taking vaccines also decreases. This indicates that individuals who are less proactive in practicing preventive measures and following recommended guidelines are less inclined to prioritize their health and well-being by getting vaccinated. Additionally, it was made clear that people refuse the H1N1 vaccine more often than the Seasonal Flu vaccine. A potential reason for is that the Seasonal Flu one is recurrent and therefore more trust is put into it, while the H1N1 is taken as a single dose. Moreover, seasonal flu spreads yearly, which could make individuals more inclined to get vaccinated against it, while the influenza virus does not have such a common outbreak nowadays. Yet should they refuse one of the vaccines, the data shows that makes them less likely to receive the other one too.

Analyzing factors that influence vaccination refusals is crucial for developing effective communication strategies and addressing fears. In turn, this leads to trust being built on healthcare practitioners and medicine in general, potentially to lives being saved and promoting better healthcare. Insights gained from this study can also be applied to other cases in the medical field, further boosting the healthcare delivery provided to the public.

As an effective communication strategy, we can suggest the demonstration of data outlining consequences of the individuals who did not get vaccinated from a new data collection, emphasizing the health benefits of taking the vaccine. Additionally, relying on credible online sources is crucial to address misinformation and build trust that will lead to individuals making an informed decision and contribution towards a higher vaccination acceptance rate.

D. Limitations

The training time for the gradient boosted algorithms has proved to be a limitation, taking several hours to find the optimal hyperparameters of each model. This is due to the extensive search space and the need to evaluate multiple combinations of parameters, leading to the extended time. However, despite the time-consuming process, the hyperparameters optimization led to improved models' performance.

We relied on self-reported data, which could've been subject to personal bias from individuals or social bias. Moreover, the dataset may not fully represent a diverse population of individuals, limiting the generalization of the results.

E. Future Improvements and Development

There are several improvements that could be followed with more time and resources. To begin with, parallel computing and other efficient algorithms such as a StackingClassifier could be explored to reduce training time and potentially increase the model's performance.

It's also important to note that achieving the lowest false positives and false negatives indicates the model's effectiveness in distinguishing between classes and making accurate predictions. However, in the future, further analysis, and evaluation of other performance metrics, such as precision, recall, and F1 score, would provide a more comprehensive understanding of the model's overall performance and its trade-offs.

In order to improve the potential performance of CatBoost, implementing bagging may enhance its performance and prevent overfitting data which is something boosting does not do. Bagging essentially on different subsets of the data, allowing it to capture diverse patterns of the data, reducing the model's sensitivity to outliers/noisy data and ultimately leading to improved predictions.

Potential further research identified from the flu shot learning dataset is the influence of social media on flu vaccine uptake. While the dataset provides insights into behaviors related to the vaccines, it lacks information on the impact of

social media. Exploring this area would involve examining the types of vaccine-related information shared on social media platforms and how it affects individuals' beliefs and attitudes towards vaccination. This investigation could uncover the role of social media in circulating misinformation or even if positive vaccine attitude is promoted. Additionally, exploring the different impacts of social media on a diverse demographic group, as for example younger individuals who heavily depend on social media as a source of health information tend to have multiple beliefs that oppose different generations, and that difference can yield valuable findings. Overall, this would contribute to the development of targeted public health campaigns with the goal of improving vaccine acceptance.

VI. CONCLUSION

In conclusion, this project explored the behavioral insights into H1N1 and seasonal vaccine refusal by using the flu shot learning dataset. In this study, we employ five different models alongside various methods of visualisation and pre-processing. By analysing the data, we uncover valuable patterns and relationships that identify factors influencing vaccine refusal alongside the implementation of diverse visualisations enhancing the reliability of our findings.

Approximately 50% of individuals have received the seasonal flu vaccine, while only 20% have been vaccinated for H1N1. Notably, there is a positive correlation of 60% between doctor recommendations for H1N1 and seasonal vaccines, indicating a likelihood of receiving both vaccines if one is administered.

However, most behavioral-based features show minimal correlation with vaccine refusal, including behaviors like taking antiviral medications, attending large gatherings, and avoidance or outside-home activities. Variables such as having a child under 6 months and household adults also do not strongly predict vaccine refusal. The data suggests that individuals who engage in behaviors like no hand washing, not wearing face masks, and frequent face-touching are less likely to receive either vaccine.

Furthermore, a gender disparity is observed, with a majority of females receiving the seasonal vaccine. This highlights the need for improved communication strategies to address barriers to vaccine acceptance among specific demographic groups.

Future research should explore the reasons behind the gender disparity in vaccine uptake and investigate the impact of doctor recommendations on vaccination behavior. Additionally, by exploring the impact of social media on vaccine uptake across different demographics, we can use targeted communication strategies for broader reach and ultimately counter the misinformation footprint on social media.

Overall, this project successfully achieved all the predefined objectives, with XGBoost achieving the high-performance accuracy of 85%, alongside providing a deeper understanding of the causations of vaccine refusal. As a result, this provides a foundation to further research that could benefit the medical field.

REFERENCES

[1] Chephra McKee (2016) Exploring the Reasons Behind Parental Refusal of Vaccines. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4869767/ (Accessed: 03 March 2023).

[2] S. S. Ayachit, T. Kumar, S. Deshpande, N. Sharma, K. Chaurasia and 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, doi: 10.1109/ICACCCN51052.2020.9362909.

[3] Inampudi, S., Johnson, G., Jhaveri, J., Niranjan, S., Chaurasia, K., Dixit, M. (2021). Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination. In: Garg, D., Wong, K., Sarangapani, J., Gupta, S.K. (eds) Advanced Computing. IACC 2020. Communications in Computer and Information Science, vol 1367. Springer, Singapore. https://doi.org/10.1007/978-981-16-0401-0_11.

[4] Zhang, R., Zhao, T. and Zhou, Y. (2021) Sign in, RPubs. Available at: https://rpubs.com/ruoyuzhang426/H1N1_Vaccine (Accessed: 14 May 2023), unpublished.

[5] Heidi J. Larson, Seth Mnookin, Chapter 27 - Trust and Confidence in Vaccines: Tales of Three Vaccines, Lessons for Others, Editor(s): Barry R. Bloom, Paul-Henri Lambert, The Vaccine Book (Second Edition), Academic Press, 2016, Pages 529-540, ISBN 9780128021743, https://doi.org/10.1016/B978-0-12-802174-3.00027-8

Contributions Table

Section	Daniel	Viki
Abstract	50%	50%
Introduction	50%	50%
Literature Review	50%	50%
Methodology	50%	50%
Results	50%	50%
Discussion	50%	50%
Conclusion	50%	50%