Predicting the Probability of H1N1 and Seasonal Influenza Vaccine Uptake using Machine Learning Methods

Song Li
Department of Computer Science
University of Nottingham
Nottingham, UK
alysl40@nottingham.ac.uk

Qi Tao Ye
Department of Computer Science
University of Nottingham
Nottingham, UK
alyqy11@nottingham.ac.uk

Abstract — **This paper analyzes the data set which is the result of the National 2009 H1N1 Flu Survey (NHFS). More than twenty thousand participants from the USA were asked questions about their basic information and opinions about H1N1 and seasonal vaccinations. The goal of this paper is to predict the rates of individuals receiving the two kinds of vaccinations. In this paper, the two authors used different data preprocessing methods, selected different features and used 5 different machine learning models to predict the likelihood of individuals receiving H1N1 and seasonal influenza vaccines. Finally, the prediction accuracy of each other is compared by calculating relevant indicators. The results show that the decision tree regression not only has high prediction accuracy but also runs fast in dealing with the problem of this high-dimensional data set.**

Keywords—Data science; Feature engineering; Machine learning; Vaccine prediction; Classification; Exploratory Data Analysis

I. Introduction

In the 2009-2010 season, NCIRD and the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC) jointly conducted NHFS which was sponsored by the National Center for Immunization and Respiratory Diseases (NCIRD). The dataset of this paper is the result of the NHFS. The NHFS was a list-assisted random-digit-dialing telephone survey of households, designed to monitor influenza immunization coverage. The number of participants is roughly about twenty thousand. The dataset contains 36 columns of features. These features describe the respondent's characteristics, habits, behaviors, and opinions about H1N1 and seasonal flu vaccines. Besides, there are training set labels dataset and test set features in addition to the 36 features dataset.

The two authors of this paper aim to predict how likely individuals are to receive their H1N1 and seasonal flu vaccines by using different approaches of machine learning. The target variables of this research are h1n1_vaccine and seasonal_vaccine which range from 0 to 1. The closer to 1, the more likely individuals will receive vaccines.

Predicting the vaccination rate of seasonal influenza vaccines in different age groups may improve the government's vaccination strategy, thereby increasing vaccination rates and reducing influenza mortality rates (Arevalo et al., 2020).

Maurer J et al. (2009) indicates that there is a correlation between the rates of H1N1 vaccination receive and seasonal vaccination receive. They also think to predict the rates of receiving both vaccinations is helpful for the government to decide how many vaccinations it should produce.

This paper predicts the rates of participants who receive H1N1 vaccination by asking about their possibility of receiving H1N1 vaccination and calculating the correlation between the rates of receiving H1N1 and seasonal vaccination and predicting based on whether they have received seasonal vaccination. The result shows that the average self-reported probability of being vaccinated against novel H1N1 was 49.6%. Importantly, the average stated probability of novel H1N1 vaccination was twice as high among vaccinated adults than among unvaccinated adults (73% vs. 34.2%) (Maurer J et al., 2009). But there is a limitation to this research, the survey of this research was conducted between May 26th and June 8th. People's intentions will change as the development of the pandemic unfolds, so the result of this paper has timeliness.

The dataset for this paper contains numerous attributes, ranging from basic participant information to their opinions towards the two vaccinations, which differs from the last research. To find the most relevant attributes, various approaches were utilized to clean the dataset. Additionally, we used and compared the performance of five different machine learning models, namely Logistic Regression, K-Nearest Neighbour Classifier(K-NN), Decision Tree Regression, Support Vector Machines(SVM) and Artificial Neural Network(ANN), by calculating the eigenvalues to compare the performance of the five models.

II. Literature Review

Vaccination is a crucial public health measure for the prevention and control of infectious diseases. By activating the immune system, vaccines enable the body to develop immunity and defend against specific pathogens. On the other hand, despite the well-established benefits of vaccination, there are still many challenges worldwide that affect the effectiveness and coverage of vaccination.

Improving vaccination rates to achieve maximum public health benefits has become a major concern for health authorities. By increasing vaccination rates, it is possible to effectively reduce the spread and outbreaks of diseases, reduce socioeconomic burdens, and protect the health and lives of the public. Furthermore, improving vaccination rates is of significant importance for public health policies and practices. Understanding the current status and influencing factors of vaccination provides a scientific basis for formulating and adjusting vaccination policies, optimizing resource allocation, coordinating healthcare services, and ensuring comprehensive coverage and effective implementation of vaccination activities.

Based on previous research, researchers have conducted extensive studies on factors influencing people's decision to receive vaccines. For example, studies have shown a positive correlation between higher education levels and vaccine uptake (Akwataghibe et al., 2019). Kennedy indicated that political ideology influences the acceptance or refusal of vaccines, with Democrats more likely to choose vaccination. Additionally, some people avoid vaccination due to religious reasons, as they may associate vaccines with Satan (Pugliese-Garcia et al., 2018).

In the 1950s to 1960s, the concept of machine learning emerged. Arthur Samuel coined the term "machine learning" in 1959. With advancements in computational power and the accumulation of data, statistical learning methods have been widely applied in machine learning. Vladimir Vapnik and Alexey Chervonenkis proposed the Support Vector Machines (SVM) method, which has achieved significant success in classification and regression problems. In addition to the SVM model, there are many other well-known machine learning models. For example, decision tree models are easy to understand and interpret, and they are efficient in handling large datasets. Random forest models can handle high-dimensional and large-scale datasets, and they can evaluate the importance of features, performing well in classification and regression problems. Neural network models simulate the connections between neurons in the human brain, allowing them to learn and discover complex nonlinear relationships.

Despite the existence of many mature models and extensive predictive research, the study of predicting vaccine uptake has not received widespread attention, particularly regarding the influence of personal experiences on vaccine uptake. Therefore, in this paper, we collected 36 personal-related features to explore the impact of personal experiences on vaccine uptake.

III. Methodology

Li and Ye used different steps and methods for data pre-processing, feature selection, modelling, model optimisation, model performance evaluation, and prediction of target variables, respectively, and their methods and steps are described next.

To predict the target variables as accurately as we can, Li uses six steps. The six steps consist of data wrangling, features selection, model selection, model training, model evaluation, and application of the model.

A. Data Wrangling

Data wrangling is a process of iterative data exploration and transformation that enables analysis. The goal of data wrangling is to make data more usable. Data wrangling plays an important role in data modeling and analysis. The result of data wrangling decides whether we can achieve the goal we wanted directly. The raw dataset consists of either too much data or too little data the most of time. Corrupt and noisy data which is invaluable for modeling and analysis should be removed to improve the accuracy. Missing values and outliers should be either replaced by other values or removed. Sometimes multiple sources of data should be combined. (Kandel S et al., 2011).

Five steps are needed in data wrangling:

- "Dataset Diagnose". Li reveals that the dataset contains many missing values and sub-typed variables, and that the data is of high dimensionality and non-uniform data type.
- Manipulation. Li replaces all of the null values with np. NaN firstly by using df. fillna function. Three variables have nearly half of the missing values, so he have decided to directly delete these three variables. They are "health_insurance", "employment_industry", and "employment_occupation". He chooses to replace missing values with mode as all of the features in the dataset are either binary or categorical data.
- Converting the numeric variables in the dataset to int and the subtype variables in the dataset to object.
- The categorical variables in the dataset need to be encoded. Li uses ordinal encoding to encode age_group, education, and income_poverty variables. The other categorical variables, such as race, sex, marital_status, etc., have an equal relationship between the different values, so he uses one-hot encoding for these variables, using the get_dummies function in the pandas library.
- Using the concat function in the pandas library to merge the training set and the target set into one dataset.

Data wrangling contains data transformation. But it is no need to normalize the dataset as most attributes are either binary or categorical data that are in the same scope. Data aggregation and enrichment are not necessarily due to the large dataset.

B. Features Selection

Although one-hot encoding can handle typed variables well, it also makes the dataset very large, and in order to improve the modelling efficiency and prediction accuracy, it is necessary to select the variables that are most correlated with the target variables to train the model. Li used the corr function from the pandas library to calculate the Pearson correlation coefficient between each variable and the target variable, and used the variables with correlation coefficients greater than 0.1 and less than negative 0.1 to train the model. 9 and 19 variables from the training set were finally selected as inputs to the model.

The final step before modelling is to divide the training set and the test set. The variables for the training model were selected based on Pearson's correlation coefficient, and the entire data set was divided using the train_test_split function, with 70 per cent of the data used for training the model and 30 per cent for testing the model performance.

C. Model Selection

The choice of model is very important, as it directly determines how well the target variable is predicted. After reviewing relevant information, Li finally chose three models, namely logistic regression model, K-NN and decision tree, to predict the target variables.

(1) Logistic Regression Model

The earliest appearance of logistic regression dates back to 1960, when Cornfield et al. first used a logistic regression model (Cornfield et al., 1960, p. 1). Walker S et al. developed a logistic regression method in 1967 to predict the probability of an event occurring in dichotomous or multichotomous data (Walker S et al., 1967). The use of logistic regression became popular in the 1980s and is now one of the most widely used research methods. Common applications of logistic regression include predicting the probability of disease and predicting customer purchase intentions (Nick TG et al., 2007).

The formula for the logistic regression model is $P(Y) = \frac{1}{1+e^{-z}}$, with z = $b0 + b1x1 + \in i$ when there is only one predictor independent variable and z = $b0 + b1x1 + b2x2 + b3x3 + \cdots + bnxn + \in i$ when there are multiple predictor independent variables (Field A, 2009). The basic assumptions of logistic regression include the absence of multicollinearity, log-linearity of continuous variables and the absence of strongly influential outliers, and independence of errors (Stoltzfus JC. 2011).

As the target variable in this study is a probability lying between 0 and 1, representing the likelihood of an individual receiving each of the two vaccines, the range of values of the target variable dictates that linear regression cannot be used but rather logistic regression, as linear regression predictions output values between (-∞,+∞), which is clearly inconsistent with the target variable.

In this thesis, the sklearn library in Python is used to invoke the logistic regression model. Two logistic regression models are built separately, and the optimal parameters of the model are found within a specific range using the grid search in the sklearn library. The accuracy of the optimal model is assessed using cross-validation in sklearn, and the average value is taken five times as the judge of the model performance.

However, there are many limitations to logistic regression, such as logistic regression being very sensitive to outliers and overfitting due to insufficient sample size or overly complex features, all of which need attention. I have considered these issues in the pre-processing stage, and I believe that the logistic regression model I have used is a good predictor of the target variable.

(2) K-Nearest Neighbour Classifier

The K-Nearest Neighbour model has been applied mainly to classification and regression problems. The first use of the K-NN model dates back to 1952, when Fix et al. (1952) first used the model to classify a sample.

The K-NN algorithm is based on calculating the Euclidean distance to find the nearest K points to each point to be classified, and then determining which class the point to be classified belongs to based on the class of these K points, and if most of these K points belong to class A, then the point to be classified also belongs to class A. The formula for calculating the Euclidean distance between points a, b is: $d(a,b) = \sqrt{\sum_{i=1}^{n}(bi - ai)^2}$. The K-NN algorithm performs well on classification and regression problems, and is widely used on binary and multi-category problems because of its ease of operation. However, the K-NN algorithm performs generally well when dealing with large, high-latitude datasets, due to the fact that the K-NN algorithm does not have true data-based learning. To improve the performance of K-NN models, it is crucial to choose the most appropriate K value (Kataria A et al., 2008). Li builds K-NN models by calling the sklearn library. to determine the best K value, the stochastic grid search method is used to build K-NN models with K values from 1 to 20 respectively and evaluate the performance. the best performing model is selected for classification, and used cross-validation to assess model performance.

(3) Decision Tree Classifier

The decision tree model is a basic classification and regression method with a tree-like structure that performs the classification task by classifying objects starting at the root node of the tree and each child node classifying the object according to one of its attributes until the final classification is completed (Quinlan J R, 1996).

Decision tree classification models have unique advantages in many ways, for example: they are good at splitting complex decision problems into simple local decisions at multiple levels, and each internal node in a decision tree regression model uses a smaller number of features to solve a high-dimensional problem. Compared to traditional classification models, the sample data in a decision tree classification model does not need to be tested for each category, thus improving computational efficiency (Rasoul S et al., 1991, p. 2).

Li invoked the decision tree classification model from the sklearn library, employed a grid search to optimise the optimal parameters of the decision tree classifier, used the optimal parameters to build the decision tree classification model, and evaluated the performance of the model through cross-validation.

D. Model Training, Model Evaluation, And Application of The Model.

Li built different classification models by calling functions in the sklearn library. To optimise the parameters of each model, he used two methods, grid search and randomized search, first defining the search range of the parameters separately, then implementing them through the GridSearchCV and

RandomizedSearchCV functions, and then evaluating the accuracy of the models through cross_val_ score to evaluate the accuracy of the models, setting the number of cross-validations to 5, and averaging the results of the 5 cross-validations to obtain the average accuracy of each model.

The pre-processed training set data was used to train different models, and the trained models were then used to predict the target variables based on the test set data.

The machine learning method adopted by Ye generally includes: data observation, data processing, feature selection, hyperparameter optimization, performance evaluation and result prediction.

A. Observation of Data

First, Ye imports the Pandas library. Next, he uses the pd.csv() method to read the data, and the obtained data is formatted as DataFrame, referred to as df. Then, use the df.info() method to observe the data, which provides information such as column names, non-null value counts, data types, and memory usage. In addition, Ye also needs to know the number of missing values in each column. This can be accomplished using the df.isna().sum() method. Furthermore, to prevent mismatched prediction results due to duplicate values, Ye uses the df_train_feature.duplicated().sum() method to check for duplicates in both the training and testing data. Finally, Ye noticed that there are many categorical features, so he examines the number of categories for each feature.

B. Data Preprocessing

After completing the data observation, Ye proceeds with data preprocessing. Firstly, for features with missing values less than 30%, he fills them with the mode. For features with missing values exceeding 30%, he deletes them since they can introduce significant errors in the prediction results. For categorical data, some features are represented as strings, but models require numeric values for prediction. Simply replacing them with natural numbers cannot adequately reflect the relative size of different categories. Therefore, he introduce one-hot encoding. This conversion can be achieved using the pd.get_dummies(df_test_feature, columns) method.

C. Feature Selection

Since the number of features exceeds 60 after applying one-hot encoding, to maintain computational efficiency and preserve prediction accuracy, Ye employs the Filter method for feature selection. Using the chi-squared test, he ranks the correlations between h1n1_vaccine and seasonal_vaccine with features and select the top 20 features with the highest correlation.

D. Hyperparameter Tuning

Next, Ye employs the Grid Search method for hyperparameter tuning. Grid Search is a method for optimizing machine learning models by systematically exploring multiple combinations of hyperparameters. Firstly, he defines the hyperparameters and their possible ranges to be tuned for both SVM and ANN. Then, a grid is created, listing all possible combinations of hyperparameters. For each combination, the model is trained on the training set and evaluated using the validation set or cross-validation. Finally, the best hyperparameter combination is selected based on metrics such as accuracy. In sklearn, the GridSearchCV module can be used to implement this process.

E. Evaluation

Lastly, Ye needs to evaluate the model's performance. He uses K-fold Cross-Validation. It allows model selection and hyperparameter tuning on a limited dataset. By averaging performance metrics over 5 iterations, he can obtain a more accurate estimate of the model's generalization ability.

F. Prediction of Results

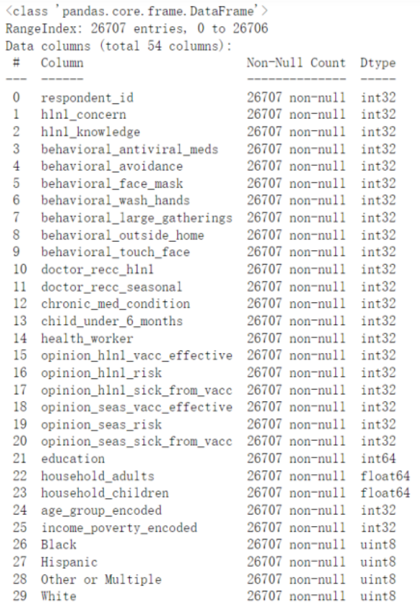
By comparing the accuracy, Ye selects the best-performing models for predicting h1n1_vaccine and seasonal_vaccine respectively. Next, he will use the training set data for prediction and export the resulting answers and probability distributions as csv files.

IV. Results from Each Of The Stages

The data processing methods used by Li and Ye are generally similar, and some details are different, but the machine learning modeling methods used by the two are quite different, which leads to the difference in the final prediction results of the two. The following will introduce the result of each stage of the two authors.

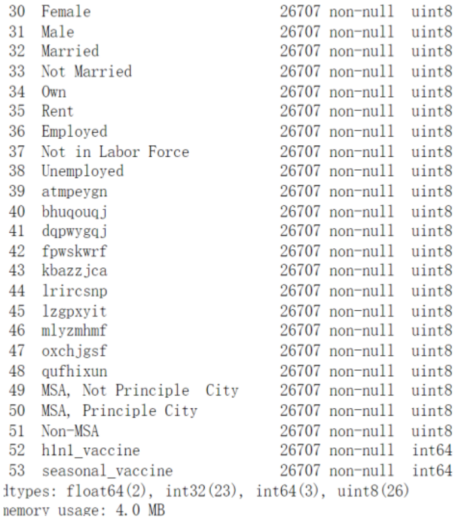
A. Li's Result

Before feature selection, Li needs to perform a series of preprocessing on the original data set, including: processing of missing values, data type conversion, categorical variable conversion, etc. The preprocessed data set has a total of 54 columns and all of them are Value type variables, some results are shown in Figure 1 and Figure 2.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26707 entries, 0 to 26706
Data columns (total 54 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   respondent_id                26707 non-null  int32
 1   h1n1_concern                 26707 non-null  int32
 2   h1n1_knowledge               26707 non-null  int32
 3   behavioral_antiviral_meds    26707 non-null  int32
 4   behavioral_avoidance         26707 non-null  int32
 5   behavioral_face_mask         26707 non-null  int32
 6   behavioral_wash_hands        26707 non-null  int32
 7   behavioral_large_gatherings  26707 non-null  int32
 8   behavioral_outside_home      26707 non-null  int32
 9   behavioral_touch_face        26707 non-null  int32
 10  doctor_recc_h1n1             26707 non-null  int32
 11  doctor_recc_seasonal         26707 non-null  int32
 12  chronic_med_condition        26707 non-null  int32
 13  child_under_6_months         26707 non-null  int32
 14  health_worker                26707 non-null  int32
 15  opinion_h1n1_vacc_effective  26707 non-null  int32
 16  opinion_h1n1_risk            26707 non-null  int32
 17  opinion_h1n1_sick_from_vacc  26707 non-null  int32
 18  opinion_seas_vacc_effective  26707 non-null  int32
 19  opinion_seas_risk            26707 non-null  int32
 20  opinion_seas_sick_from_vacc  26707 non-null  int32
 21  education                    26707 non-null  int64
 22  household_adults             26707 non-null  float64
 23  household_children           26707 non-null  float64
 24  age_group_encoded            26707 non-null  int32
 25  income_poverty_encoded       26707 non-null  int32
 26  Black                        26707 non-null  uint8
 27  Hispanic                     26707 non-null  uint8
 28  Other or Multiple            26707 non-null  uint8
 29  White                        26707 non-null  uint8
```

Figure 1. Processed data information



```
 30  Female                        26707 non-null  uint8
 31  Male                          26707 non-null  uint8
 32  Married                       26707 non-null  uint8
 33  Not Married                   26707 non-null  uint8
 34  Own                           26707 non-null  uint8
 35  Rent                          26707 non-null  uint8
 36  Employed                      26707 non-null  uint8
 37  Not in Labor Force            26707 non-null  uint8
 38  Unemployed                    26707 non-null  uint8
 39  atmpeygn                      26707 non-null  uint8
 40  bhuqouqj                      26707 non-null  uint8
 41  dqpwygqj                      26707 non-null  uint8
 42  fpwskwrf                      26707 non-null  uint8
 43  kbazzjca                      26707 non-null  uint8
 44  lrircsnp                      26707 non-null  uint8
 45  lzgpxyit                      26707 non-null  uint8
 46  mlyzmhmf                      26707 non-null  uint8
 47  oxchjgsf                      26707 non-null  uint8
 48  qufhixun                      26707 non-null  uint8
 49  MSA, Not Principle  City      26707 non-null  uint8
 50  MSA, Principle City           26707 non-null  uint8
 51  Non-MSA                       26707 non-null  uint8
 52  h1n1_vaccine                  26707 non-null  int64
 53  seasonal_vaccine              26707 non-null  int64
dtypes: float64(2), int32(23), int64(3), uint8(26)
memory usage: 4.0 MB
```

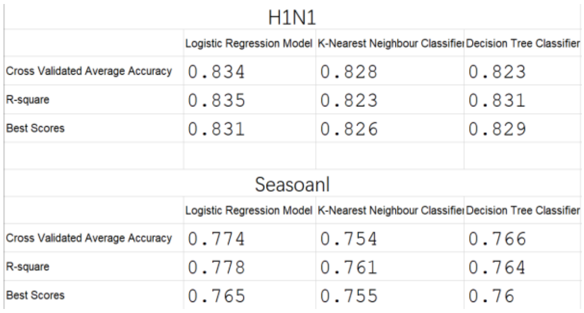
Figure 2. Processed data information

Variables with high correlation are selected for training the model by calculating the Pearson correlation coefficient between each independent variable and the target variable. Select the variables whose correlation is greater than 0.1 and less than -0.1. The results show that there are 9 variables used to predict the H1N1 vaccination rate, which are: 'h1n1_concern', 'h1n1_knowledge', 'doctor_recc_h1n1', 'doctor_recc_seasonal', 'health_worker','opinion_h1n1_vacc_effective','opinion_h1n1_risk', 'opinion_seas_vacc_effective', 'opinion_seas_risk'， There are 19 variables used to predict the Seasonal vaccination rate, namely:'h1n1_concern','h1n1_knowledge','behavioral_wash_hands','behavioral_touch_face','doctor_recc_h1n1','doctor_recc_seasonal','chronic_med_condition','health_worker','opinion_h1n1_vacc_effective','opinion_h1n1_risk','opinion_seas_vacc_effective','opinion_seas_risk','household_children','age_group_encoded', 'White', 'Own', 'Rent', 'Employed', 'Not in Labor Force'.

After preprocessing and feature screening of the database, the qualified variables are used to train the model respectively. The accuracy and running time of different models will be presented below.

(1) Accuracy

Cross-validate the three models separately, set the number of times of cross-validation to 5, and calculate the average accuracy rate of 5 times, and calculate the R-square and grid search on the given test data for the three models. The score of the best model of , and compare the accuracy of the three models through these three performance indicators. The result is shown in Figure 3.



H1N1

	Logistic Regression Model	K-Nearest Neighbour Classifie	Decision Tree Classifier
Cross Validated Average Accuracy	0.834	0.828	0.823
R-square	0.835	0.823	0.831
Best Scores	0.831	0.826	0.829

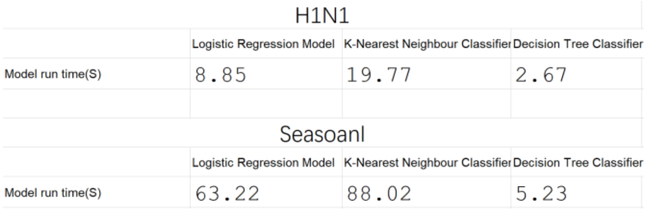
Seasoanl

	Logistic Regression Model	K-Nearest Neighbour Classifie	Decision Tree Classifier
Cross Validated Average Accuracy	0.774	0.754	0.766
R-square	0.778	0.761	0.764
Best Scores	0.765	0.755	0.76

Figure 3: Accuaracy

(2) Running time

When comparing the performance of different model algorithms, the average running time of the model is an important evaluation indicator. The faster the running time of the model can not only save costs but also improve efficiency. By using the time module to calculate the time from the beginning of training to predicting the target variable for each model, and taking multiple measurements and averaging, the results are shown in Figure 3.



H1N1

	Logistic Regression Model	K-Nearest Neighbour Classifier	Decision Tree Classifier
Model run time(S)	8.85	19.77	2.67

Seasoanl

	Logistic Regression Model	K-Nearest Neighbour Classifier	Decision Tree Classifier
Model run time(S)	63.22	88.02	5.23

Figure 4: Running time

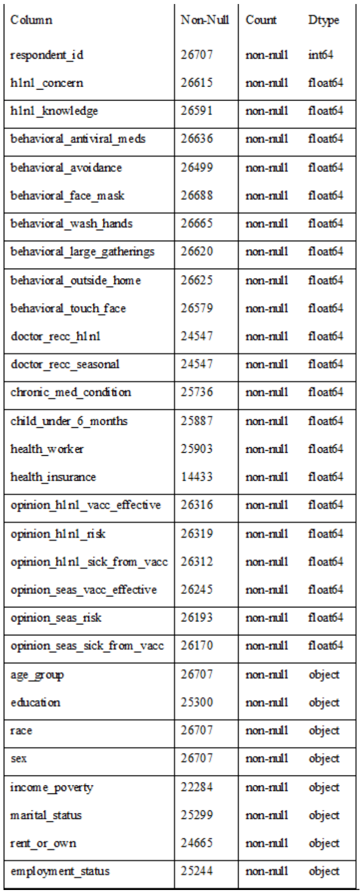
B. Ye's Result

The following introduces the results of Ye data observation, data preprocessing, feature selection, hyperparameter optimization, and performance evaluation.

The following introduces the results of Ye data observation, data preprocessing, feature selection, hyperparameter optimization, and performance evaluation.

(1) Observation of Data:

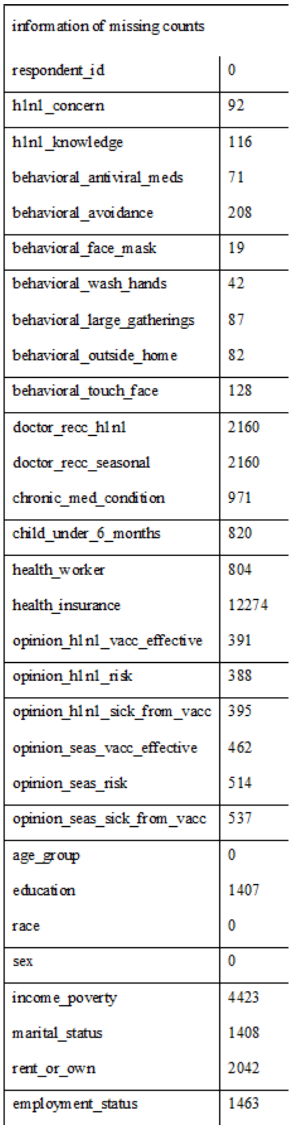
The dataset consists of 26,707 rows and 36 features. Among them, 24 features are numerical, while the remaining 12 features are in string format. Except for the primary key "respondent_id," every column has missing values.



Column	Non-Null	Count	Dtype
respondent_id	26707	non-null	int64
h1n1_concern	26615	non-null	float64
h1n1_knowledge	26591	non-null	float64
behavioral_antiviral_meds	26636	non-null	float64
behavioral_avoidance	26499	non-null	float64
behavioral_face_mask	26688	non-null	float64
behavioral_wash_hands	26665	non-null	float64
behavioral_large_gatherings	26620	non-null	float64
behavioral_outside_home	26625	non-null	float64
behavioral_touch_face	26579	non-null	float64
doctor_recc_h1n1	24547	non-null	float64
doctor_recc_seasonal	24547	non-null	float64
chronic_med_condition	25736	non-null	float64
child_under_6_months	25887	non-null	float64
health_worker	25903	non-null	float64
health_insurance	14433	non-null	float64
opinion_h1n1_vacc_effective	26316	non-null	float64
opinion_h1n1_risk	26319	non-null	float64
opinion_h1n1_sick_from_vacc	26312	non-null	float64
opinion_seas_vacc_effective	26245	non-null	float64
opinion_seas_risk	26193	non-null	float64
opinion_seas_sick_from_vacc	26170	non-null	float64
age_group	26707	non-null	object
education	25300	non-null	object
race	26707	non-null	object
sex	26707	non-null	object
income_poverty	22284	non-null	object
marital_status	25299	non-null	object
rent_or_own	24665	non-null	object
employment_status	25244	non-null	object

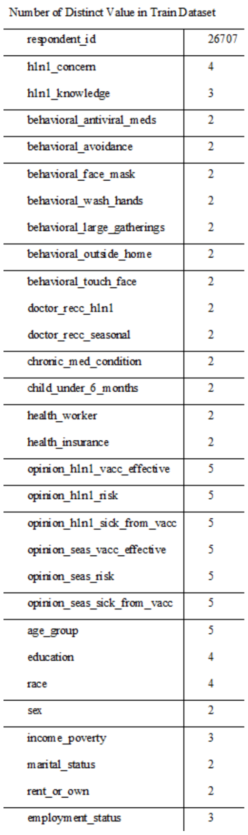
Figure 5 Information of Data

Regarding the missing values, most features have less than 5000 missing entries, except for the 'health_insurance', 'employment_industry', and 'employment_occupation' columns, which have a significantly higher number of missing values. They exceed the defined threshold of 30% for missing values. Furthermore, all features are categorical and there are no duplicate entries.



information of missing counts	
respondent_id	0
h1n1_concern	92
h1n1_knowledge	116
behavioral_antiviral_meds	71
behavioral_avoidance	208
behavioral_face_mask	19
behavioral_wash_hands	42
behavioral_large_gatherings	87
behavioral_outside_home	82
behavioral_touch_face	128
doctor_recc_h1n1	2160
doctor_recc_seasonal	2160
chronic_med_condition	971
child_under_6_months	820
health_worker	804
health_insurance	12274
opinion_h1n1_vacc_effective	391
opinion_h1n1_risk	388
opinion_h1n1_sick_from_vacc	395
opinion_seas_vacc_effective	462
opinion_seas_risk	514
opinion_seas_sick_from_vacc	537
age_group	0
education	1407
race	0
sex	0
income_poverty	4423
marital_status	1408
rent_or_own	2042
employment_status	1463

Figure 6 Information of Missing Counts



Number of Distinct Value in Train Dataset

respondent_id	26707
h1n1_concern	4
h1n1_knowledge	3
behavioral_antiviral_meds	2
behavioral_avoidance	2
behavioral_face_mask	2
behavioral_wash_hands	2
behavioral_large_gatherings	2
behavioral_outside_home	2
behavioral_touch_face	2
doctor_recc_h1n1	2
doctor_recc_seasonal	2
chronic_med_condition	2
child_under_6_months	2
health_worker	2
health_insurance	2
opinion_h1n1_vacc_effective	5
opinion_h1n1_risk	5
opinion_h1n1_sick_from_vacc	5
opinion_seas_vacc_effective	5
opinion_seas_risk	5
opinion_seas_sick_from_vacc	5
age_group	5
education	4
race	4
sex	2
income_poverty	3
marital_status	2
rent_or_own	2
employment_status	3

Figure 7 Number of Distinct Value in Train Dataset

(2) Data Preprocessing

The data with excessive missing values was removed, and the remaining features that are categorical in string format were encoded using one-hot encoding. As a result, the dataset now has 26,707 rows and 61 features.

(3) Feature Selecting

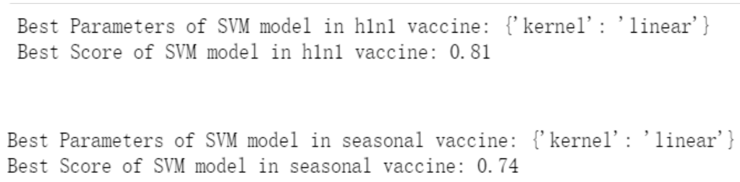
After ranking the model's correlation coefficients for predicting H1N1 vaccine and seasonal vaccine, Ye selected the top 20 features. They are as follows:

H1N1vaccine:respondent_id,h1n1_concern,h1n1_knowledge,behavioral_antiviral_meds,behavioral_face_mask,behavioral_touch_face,doctor_recc_h1n1,doctor_recc_seasonal,chronic_med_condition,child_under_6_months,health_worker,opinion_h1n1_vacc_effective,opinion_h1n1_risk,opinion_h1n1_sick_from_vacc,opinion_seas_vacc_effective,opinion_seas_risk,education_CollegeGraduate,race_Black,income_poverty_> $75,000, hhs_geo_region_bhuqouqj.

Seasonalvaccine:respondent_id,h1n1_concern,h1n1_knowledge,behavioral_touch_face,doctor_recc_h1n1,doctor_recc_seasonal,chronic_med_condition,health_worker,opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_seas_vacc_effective, opinion_seas_risk, household_children, age_group_18 - 34 Years, age_group_35 - 44 Years, age_group_65+ Years, rent_or_own_Rent,employment_status_Employed,employment_status_Not in Labor Force, employment_status_Unemployed.

(4) Hyperparameters Tuning

The optimal hyperparameters obtained by SVM using grid search method are as follow:

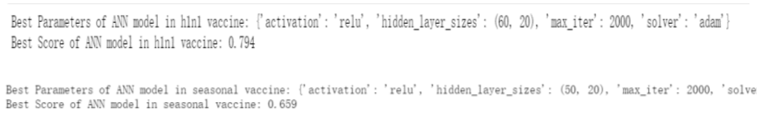


```
Best Parameters of SVM model in h1n1 vaccine: {'kernel': 'linear'}
Best Score of SVM model in h1n1 vaccine: 0.81


Best Parameters of SVM model in seasonal vaccine: {'kernel': 'linear'}
Best Score of SVM model in seasonal vaccine: 0.74
```

Figure8 Best Parameters of SVM model

The optimal hyperparameters obtained by ANN using grid search method are as follow:



```
Best Parameters of ANN model in h1n1 vaccine: {'activation': 'relu', 'hidden_layer_sizes': (60, 20), 'max_iter': 2000, 'solver': 'adam'}
Best Score of ANN model in h1n1 vaccine: 0.794


Best Parameters of ANN model in seasonal vaccine: {'activation': 'relu', 'hidden_layer_sizes': (50, 20), 'max_iter': 2000, 'solve
Best Score of ANN model in seasonal vaccine: 0.659
```

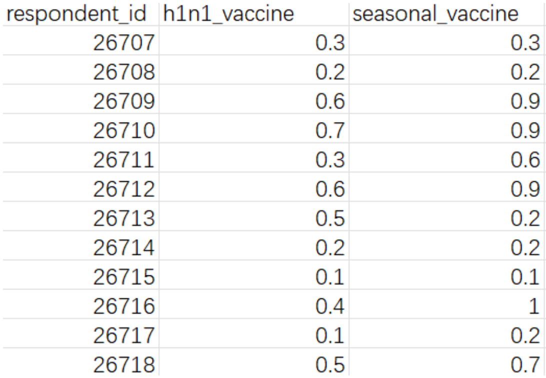
Figure9 Best Parameters of ANN model

(5) Evaluation

For SVM, Ye uses K-fold Cross-Validation to get the accuracy of prediction for h1n1 vaccine, the accuracy is 82.37% in train dataset and 66.82% in test dataset. With regard to seasonal vaccine, the accuracy is 76.88% in train dataset and 54.29% in test dataset.

For ANN, Ye uses K-fold Cross-Validation to get the accuracy of prediction for h1n1 vaccine, the accuracy is 87.56% in train dataset and 67.21% in test dataset. With regard to seasonal vaccine, the accuracy is 83.22% in train dataset and 51.13% in test dataset.

Then, Ye made predictions on the vaccination status of the test data and exported the file as follows:



respondent_id	h1n1_vaccine	seasonal_vaccine
26707	0.3	0.3
26708	0.2	0.2
26709	0.6	0.9
26710	0.7	0.9
26711	0.3	0.6
26712	0.6	0.9
26713	0.5	0.2
26714	0.2	0.2
26715	0.1	0.1
26716	0.4	1
26717	0.1	0.2
26718	0.5	0.7

Figure10 Screenshot of predicted results

V. Discussion

In this study, Li established a logistic regression model, a K-NN model, and a decision tree regression model to predict the probability of individuals being vaccinated against H1N1 and Seasonal vaccines, respectively.

Firstly, to obtain the optimal logistic regression model, find the model with the highest score through grid search and cross-validation, and input the preprocessed training data into the logistic regression model. The results show that: the logistic regression model predicts individual vaccination with the accuracy rate of the H1N1 vaccine is 83%, and the accuracy rate of Seasonal vaccination is about 77%. The running time of the model of inoculation rate is 63 seconds, which shows that compared with the other two models, the logistic regression model has certain advantages in dealing with high-dimensional data sets, but it is worth noting that as the dimension increases, the running time of the model Time will be very slow.

Then, the above experimental results reflect that the accuracy of the K-NN model in predicting the personal vaccination rate is between 75% and 82%, but the running time is longer, because the K-NN model is performing a process on each data point. When classifying, it is necessary to repeatedly calculate the distance between the point and the surrounding points, which causes a lot of unnecessary calculations and reduces the efficiency of the model operation.

The decision tree model is good at using a small number of features to solve high-dimensional problems. The model does not need to be tested for each category, which greatly improves the operating efficiency. The accuracy of the decision tree model in predicting the probability of personal vaccination of H1N1 vaccine and Seasonal vaccine are 82% and 76%. The running time of the decision tree regression model is 2.67 seconds and 5.23 seconds respectively, which is very fast .

Through this study, Ye compared the performance of two machine learning algorithms which are SVM and ANN in predicting the administration of h1n1 vaccine and seasonal vaccine.

Firstly, Ye analyzed SVM. On the training dataset, SVM demonstrated high accuracy in predicting h1n1 vaccine and seasonal vaccine, reaching 82.37% and 76.88% respectively. However, when applying the model to the test dataset, the accuracy dropped to 66.82% and 54.29%. This indicates that SVM may suffer from overfitting when faced with unseen data. Additionally, Ye observed that SVM had a relatively slow speed in adjusting hyperparameters, especially when dealing with large-scale datasets. Therefore, despite SVM's excellent accuracy, there are some limitations in terms of efficiency.

Next, Ye examined the performance of ANN. On the training dataset, ANN achieved high levels of accuracy in predicting h1n1 vaccine and seasonal vaccine, reaching 87.56% and 83.22% respectively. However, on the test dataset, the accuracy dropped to 67.21% and 51.13%. This suggests that ANN may also face overfitting issues when encountering unseen data. On the other hand, ANN demonstrated a fast speed in adjusting weights and learning process, which is advantageous for handling large-scale datasets.

Above all, the predictions for H1N1 vaccine administration achieved a high level of accuracy, which can provide valuable insights for public health departments to understand the factors influencing vaccine uptake. This information can be instrumental in formulating and adjusting vaccine administration policies, optimizing resource allocation, coordinating healthcare services, and ensuring comprehensive coverage and effective implementation of vaccination activities.

However, the prediction for seasonal vaccine administration did not yield satisfactory results. This can be attributed to two main factors: the potential lack of consideration for class imbalance and the complexity of the employed SVM and ANN models leading to overfitting.

VI. Conclusions And Recommendation

Vaccination is a key measure for preventing and controlling infectious diseases, but we face various challenges on a global scale. Increasing vaccination rates has become a priority for health authorities to achieve maximum public health benefits. However, getting people vaccinated is not a simple task due to the complex interplay of multiple factors.

In this regard, machine learning techniques have shown promise in helping us predict the likelihood of individual vaccination. By employing machine learning algorithms, we can utilize existing data to build models that predict whether individuals are willing to receive specific vaccines. These models can take into account various factors such as personal characteristics, social factors, and psychological factors, thus providing more accurate predictions.

In summary, by considering the accuracy and running time of the five different classification models we can conclude that the decision tree classification model is highly accurate and runs the fastest in predicting individual vaccination with H1N1 and seasonal vaccines, and therefore we believe that of the five models, the decision tree model is the best in dealing with this problem.

However, we must also be aware of the limitations of machine learning models. The predictive results of the models may be influenced by data limitations and biases, as these data may not fully reflect the complexities of the real world. Additionally, the predictive performance of machine learning models still requires improvement, necessitating better handling of data.

In the future, further research and practice are needed to enhance the accuracy and reliability of machine learning models. We need to collect more data and ensure the quality and diversity of the data to make the models more comprehensive and accurate

VII. References

[1] Arevalo P, McLean H Q, Belongia E A, et al. Earliest infections predict the age distribution of seasonal influenza A cases[J]. Elife, 2020, 9: e50060.

[2] Akwataghibe, Ngozi N., et al. "Exploring factors influencing immunization utilization in Nigeria—a mixed methods study." Frontiers in public health 7 (2019): 392.

[3] Cornfield J, Gordon T, Smith W W. Quantal response curves for experimentally uncontrolled variables[J]. Bull Int Stat Inst, 1961, 38(3): 97-115.

[4] Fix, Evelyn, and Joseph L. Hodges Jr. Discriminatory analysis-nonparametric discrimination: Small sample performance. California Univ Berkeley, 1952.

[5] Field A. Logistic regression[J]. Discovering statistics using SPSS, 2009, 264: 315.

[6] Kandel S, Heer J, Plaisant C, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data[J]. Information Visualization, 2011, 10(4): 271-288.

[7] Kennedy, Jonathan. "Populist politics and vaccine hesitancy in Western Europe: an analysis of national-level data." European journal of public health 29.3 (2019): 512-516.

[8] Kataria A, Singh M D. A review of data classification using k-nearest neighbour algorithm[J]. International Journal of Emerging Technology and Advanced Engineering, 2013, 3(6): 354-360.

[9] Maurer J, Harris K M, Parker A, et al. Does receipt of seasonal influenza vaccine predict intention to receive novel H1N1 vaccine: evidence from a nationally representative survey of US adults[J]. Vaccine, 2009, 27(42): 5732-5734.

[10] Nick T G, Campbell K M. Logistic regression[J]. Topics in biostatistics, 2007: 273-301.

[11] Pugliese-Garcia, Miguel, et al. "Factors influencing vaccine acceptance and hesitancy in three informal settlements in Lusaka, Zambia." Vaccine 36.37 (2018): 5617-5624.

[12] Quinlan J R. Learning decision tree classifiers[J]. ACM Computing Surveys (CSUR), 1996, 28(1): 71-72.

[13] Stoltzfus J C. Logistic regression: a brief primer[J]. Academic emergency medicine, 2011, 18(10): 1099-1104.

[14] Walker S H, Duncan D B. Estimation of the probability of an event as a function of several independent variables[J]. Biometrika, 1967, 54(1-2): 167-179.

VIII. Contribution Section

Li's section: Abstract, Introduction, Consolidation and Formatting of the paper.

Ye’s section: Literature review

All sections involved: Methodology, Results, Discussion, Conclusion.