Machine Learning-Based H1N1 and Seasonal Flu Prediction

Syed Bukhari
University of Nottingham
MSc Computer Science
Dr. Helena Webb

Machine Learning-Based H1N1 and Seasonal Flu Prediction

An extensive data-driven examination of the correlation between predictive models and pupil's background

Syed Bukhari

University of Nottingham

MSc Computer Science

Dr. Helena Webb

Thesis Presentation Outline

Introduction

01

Background

The study explored machine learning to predict vaccination likelihood, focusing on demographics and comparing methods to improve accuracy.

02

Need for Research

Close to a million deaths. Demographics and behaviours influence both aid in public health strategies. Comparing methods improves prediction accuracy for better prevention.

The Problem We Solve

Problem And Who Has It?

Exploring current situation

Low H1N1 and flu vaccination rates in vulnerable groups raise outbreak risks, strain healthcare, and increase costs. Mistrust, misinformation, and hesitancy worsen the problem, leading to preventable deaths, economic strain, and ineffective public health responses.

Ask yourself

Why is this a problem? Current models overlook complexity of vaccine hesitancy (for study's relating to the 2009 data set). This wastes resources and leads to low vaccination rates. This sustains outbreaks, worsens health inequality, and strains healthcare systems.

Explanation

The problem we are solving is building accurate machine learning models to predict vaccination uptake, improving targeting for public health by identifying key factors influencing vaccine decisions.

Aims & Objectives

Research Objectives

Model Development

To develop a machine learning model that predicts likelihood of individuals receiving vaccinations for H1N1 and seasonal flu.

Demographic Influence

To identify key demographic factors influencing vaccination uptake.

Methodology Comparison

Compare the dual methodological approaches to improve model accuracy and generalisability.

Research Questions

01

RQ #1

How do demographic factors (age, education, income) influence perceptions of H1N1 and seasonal flu vaccine effectiveness and risks?

02

RQ #2

Is there a correlation between engaging in preventative health behaviours (hand-washing, wearing masks) and the likelihood of getting vaccinated?

03

RQ #3

How do dual methodological approaches compare at different stages of the research process?

Literature review

Theory 01

Hussain and Fatima (2020) achieved 87% accuracy with Random Forest, excluding Naïve Bayes. Ayachit et al. (2020) used CatBoost, achieving 86.17% accuracy for flu vaccination, avoiding less effective models.

Theory 02

Venkatramanan et al. (2021) used anonymized mobility data for influenza forecasting, highlighting important of real-time data. Venna et al. (2019) applied LSTM models to improve flu predictions using environmental data.

Theory 03

Byrne et al. (2011) used Health Belief Model to predict vaccine uptake based on behavioural and psychological factors.

Theory 04

Identified weakness in literature review: current models are challenged by issues such as data quality and generalisability of models across regions. Models are limited by reliance on single-method approaches.

Methodology - General

A **quantitative study** using machine learning and statistical analysis to predict vaccination likelihood for H1N1 and seasonal flu.

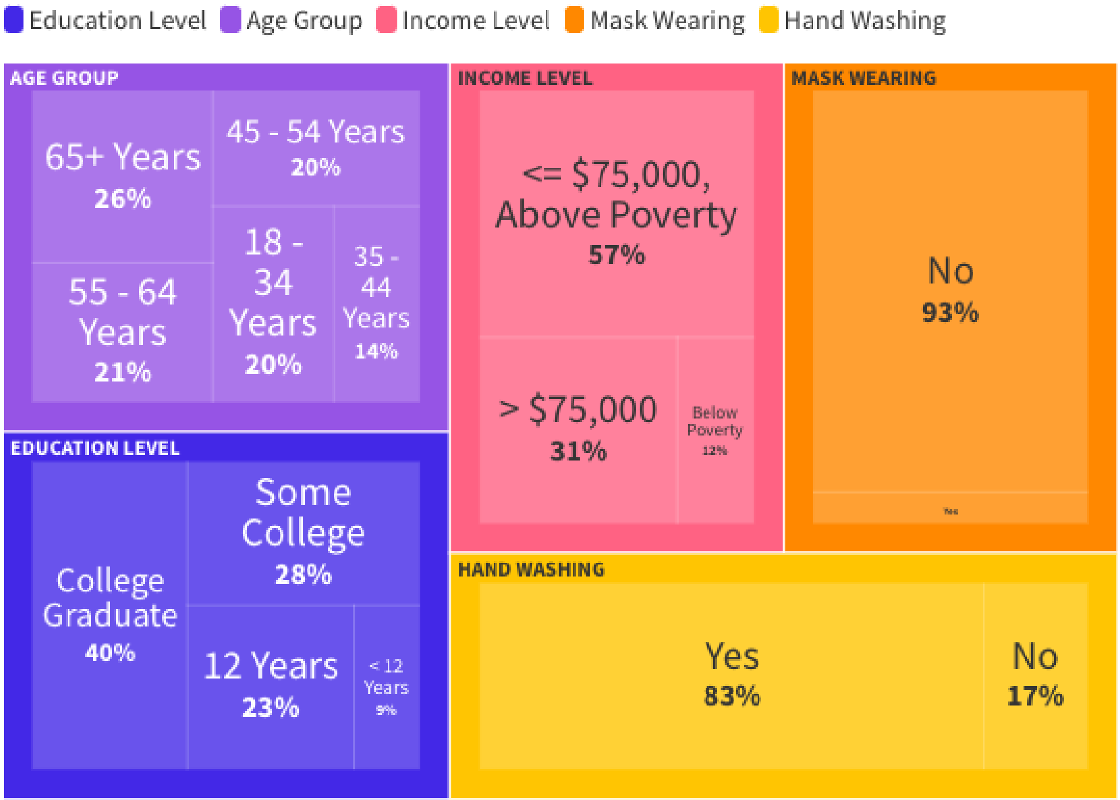
Multivariate logistic regression was used to analyse the impact of multiple demographic factors on vaccination status.

Data from **National 2009 H1N1 Flu Survey** (NHFS) of over **53,416** respondents was used. (26,708 training; 26,708 testing)

MLR for RQ1 and RQ2. **Different** preprocessing, feature selection, hyperparameter tuning, model training & evaluation applied to **Method A and B** for RQ3.

Group of Study

Demographic and Behavioural Factors Influencing Vaccination Perceptions



Methodology (RQ1 & RQ2)

01

In MLR, an interaction term multiplies two independent variables to show their combined effect on the dependent variable, beyond each variable's individual impact. For example, it reveals how hand-washing and mask-wearing together affect vaccination likelihood, while still considering their separate influences, uncovering relationships not visible when examined individually.

02

Positive values (e.g., 0.247, 0.419, 0.269): These indicate a positive relationship between factors and the outcome (e.g., vaccine uptake or perception of effectiveness/risk). A higher positive number suggests an increased likelihood or stronger belief in the outcome.

Negative values (e.g., -0.362, -0.129, -0.224): These indicate a negative relationship, meaning a lower or more negative number suggests a decreased likelihood or weaker belief in the outcome.

Research Results #1

RQ1: How do demographics impact perceptions of H1N1 and flu risks?

Age Group:

Younger adults (18-34) perceive H1N1 vaccine positively (0.247) with neutral risk (-0.029), while older adults (65+) view it as less effective (-0.005) and risker (-0362).

Younger adults moderately trust the seasonal flu vaccine (0.054) but see it as riskier (-0.129), while older adults strongly believe in its effectiveness (0.484) and perceive minimal risk (0.057).

Education:

College graduates trust H1N1 vaccine's effectiveness (0.269) with lower perceived risk (-0.224), while those with less education are skeptical (-0.018) and perceive higher risk (0.117)

For seasonal flu vaccine, college graduates show strong belief in its effectiveness (0.330) with low risk (-0.115), while those with less education are moderately trusting (0.065) but perceive higher risk (0.058).

Income:

Higher-income individuals (>$75,000) perceive H1N1 vaccine as highly effective (0.419) with lower risks (-0.180), while lower-income individuals are less confident (0.062) and perceive it as riskier (0.092).

Similarly, higher-income individuals trust the seasonal flu vaccine's effectiveness (0.534) and see low risk (-0.019), while lower-income groups are less confident (0.050) and perceive higher risk (0.068).

Research Results #2

RQ2: Does engaging in preventative behaviours correlate with vaccination likelihood?

01

Hand-Washing Frequency

Coefficient: H1N1 – 0.4401; Seasonal Flu – 0.5394; Frequent hand-washing positively correlates with higher likelihood of vaccination.

02

Face Mask Usage

Coefficient: H1N1 – 0.7440; Seasonal Flu – 0.4034; Wearing face masks shows a strong positive effect on vaccination likelihood.

03

Antiviral Medication

Coefficient: H1N1 – 0.4341; Seasonal Flu – 0.37668; Use of antiviral medication is positively associated with vaccination.

Methodology (RQ3)

Methodology A and B used different tools and techniques at each stage

Preprocessing

- **Method A**: No duplicates found. Missing data was handled using Iterative Imputation to maintain relationships between variables.

- **Method B**: No duplicates found. Mean and mode imputation were used for missing values to ensure statistical consistency.

Feature Selection

- **Method A**: Exploratory Data Analysis aggregated behavioural factors like “cleanliness” and “opinion” metrics, simplifying analysis and enhancing model interpretability.

- **Model B**: SelectKBest with chi-squared test identifying top 10 significant features for each target (H1N1 and seasonal flu), reducing dimensionality and improving model focus.

Methodology (RQ3)

Methodology A and B used different tools and techniques at each stage

Hyperparameter Tuning

- **Method A**: GridSearchCV and RandomisedSearchCV were employed to exhaustively explore hyperparameters for applied machine learning models, ensuring best configuration for model accuracy.
- **Method B**: HalvingGridSearchCV efficiently narrow down optimal parameters, balancing resource efficiency and model performance

Model Training & Evaluation

- **Method A:** Models were trained using GridSearchCV with ROC-AUC and cross validation as evaluation metrics, achieving strong results but with risk of overfitting.
- **Method B**: Applied stratified k-fold cross-validation, focusing on model accuracy and ensuring generlisation across different data subsets for real-world reliability.

Research Results #3

RQ3: How do dual-method approaches perform across different stages of flu prediction?

Preprocessing

Method A

+ High data quality (thorough cleaning and outlier handling)

- Resource Intensive (computationally extensive)

Method B

+ Fast preprocessing (mean and mode imputation speeds up preparation)

- Potentially misses subtle patterns (simpler methods may overlook nuanced data relationships)

Feature Selection & Engineering

Method A

+ High model accuracy (rich feature set captures complex relationships)

- Time consuming (EDA and feature reduction require significant effort)

Method B

+ Efficient (Simpler, faster feature selection using SelectKBest)

- Potential feature loss (May discard important, less obvious features)

Hyperparameter Tuning

Method A

+ Highly optimized models (Exhaustive search fine-tunes performance)

- Overfitting risk (Higher chance of overfitting due to extensive tuning)

Method B

+ Efficient tuning (HalvingGridSearchCV reduces overfitting and resource use)

- Lower peak accuracy (does not achieve same high accuracy as exhaustive tuning)

Model Training & Evaluation

Method A

+ High in-simple accuracy (ensemble methods capture complex data)

- Overfitting (Cross-val performance drops, limiting real-world use)

Method B

+ Generalisability (Consistent performance across data)

- Slightly lower accuracy (Peak performance is lower compared to Methodology A)

Research Results #3

Classifier	Tuning Method	Training Accuracy	Cross-Validation Accuracy
RandomForestClassifier	Grid Search	91.76%	83.92%
	Randomized Search	98.29%	84.24%
XGBClassifier	Grid Search Hyperparameter Tuning	87.18%	84.15%
	Randomized Search	94.79%	83.70%
CatBoostClassifier	Randomized Search	92.02%	85.00%
	Default Parameters	94.34%	85.26%

Rank	Model	Training H1N1 Accuracy	Test H1N1 Accuracy	Training Seasonal Accuracy	Test Seasonal Accuracy	Stratified CV H1N1 Accuracy	Stratifie CV Seasonal Accuracy
4	LogisticRegression	83.24%	84.86%	77.53%	74.21%	82.96%	77.53%
5	RidgeClassifierCV	83.33%	84.30%	77.11%	74.02%	82.96%	76.73%
10	LDA	83.10%	84.30%	77.29%	73.83%	82.91%	76.78%
1	XGBClassifier	90.03%	84.30%	86.66%	75.14%	82.63%	77.99%
9	Decision Tree	83.66%	83.55%	78.09%	73.46%	82.63%	77.62%
8	SVC	83.57%	84.49%	80.15%	75.33%	82.63%	77.48%
0	Random Forest	83.80%	83.93%	78.14%	73.64%	82.44%	77.20%
3	AdaBoost	83.19%	84.67%	77.06%	74.02%	82.40%	77.53%
2	GradientBoosting	84.74%	84.30%	80.99%	74.58%	81.88%	77.90%
7	KNN	83.99%	83.18%	82.30%	74.21%	81.60%	76.73%
6	GaussianNB	80.15%	82.99%	74.72%	72.90%	79.77%	73.97%

Discussion

Discussion #1

Results: Demographic factors like education and income significantly influence vaccine perceptions, and behaviours like hand-washing and mask-wearing increase vaccination likelihood. This guides public health interventions.

Discussion #2

Major Findings: Higher education and income levels improve vaccine uptake, and combined preventative behaviours significantly boost vaccination likelihood.

Discussion #3

Relation to past findings: Similar results in relation to accuracy scores. However, this is fresh specific questions relating to the 2009 H1N1 and Seasonal Flu data set – with inclusion of combined behaviours.

Discussion #4

Surprising Results: Method A, though more accurate, overfitted the data, while Method B, focused on efficiency, provided more generalisable, real-world results.

Suggestions & Recommendations

Recommendation

Merge Method A's accuracy with Method B's efficiency, as we now understand what works best with this dataset – done by combining tools and techniques.

Recommendation

Apply Boosted Random Forest to (potentially) further improve accuracy in predicting vaccine uptake, and evaluate models using precision, recall, and F1-score, for further deeper insights.

Suggestion #1

Incorporate more recent datasets, including COVID-19 vaccination data, to validate if similar patterns exist between demographic factors and vaccine uptake.

Suggestion #2

Use more granular age brackets and behavioural segmentation to better understand specific patterns, allowing for more tailored interventions in public health campaigns.

Suggestion #3

Future work could account for vaccine hesitancy driven by religious and cultural concerns, such as Muslim or vegetarian objections to ingredients like gelatin in vaccination which may affect uptake in certain groups.

Limitations of the Study

01

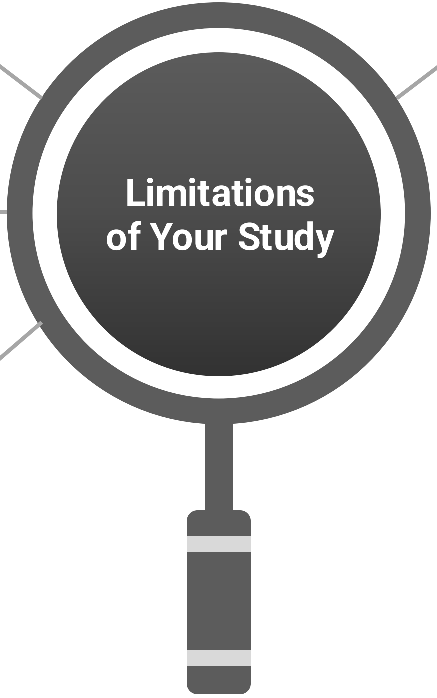
Data Set Recency
The 2009 dataset limits relevance to current vaccines.

02

Overfitting
Method A risks overfitting, reducing generalizability

03

Broad Demographics
Broader demographics could reveal more specific trends between cultures and ethnic backgrounds.



04

High Computation
Ensemble methods and tuning required extensive computational resources.

References

1. **Exploring Machine Learning Utilization on Influenza Pandemic Dataset**
   Hussain, S. and Fatima, U. (2024).

2. **Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning approach**
   Ayachit, S.S., Kumar, T., Deshpande, S., Sharma, N., Chaurasia, K. and Dixit, M. (2020).

3. **Predicting intention to uptake H1N1 influenza vaccine in a university sample**
   Byrne, C., Walsh, J., Kola, S. and Sarma, K.M. (2011).

4. **Forecasting influenza activity using machine-learned mobility map**
   Venkatramanan, S., Sadilek, A., Fadikar, A., Barrett, C.L., Biggerstaff, M., Chen, J., Dotiwalla, X., Eastham, P., Gipson, B., Higdon, D., Kucuktunc, O., Lieber, A., Lewis, B.L., Reynolds, Z., Vullikanti, A.K., Wang, L. and Marathe, M. (2021).

5. **A Novel Data-Driven Model for Real-Time Influenza Forecasting**
   Venna, S.R., Tavanaei, A., Gottumukkala, R.N., Raghavan, V.V., Maida, A.S. and Nichols, S. (2019).

Conclusions / Findings

1. How do demographics influence vaccine perceptions?

2. How do preventative behaviours impact vaccine uptake?

3. How effective are different preprocessing and modelling methods?

1. Demographics (e.g. education, income) affect vaccine views.

2. Preventative behaviours like hand washing increase vaccine uptake.

3. Method A had higher accuracy but overfitted; Method B was more generalisable.

Explores unexplored area by analysing demographic factors and preventative behaviours impact vaccine uptake using this specific data set.

More computational power could have saved time and allocated testing additional tools and techniques.

Use recent data.

Refine models by combining both methods.

QUESTIONS

THANK YOU !

Machine Learning-Based H1N1 and Seasonal Flu Prediction

Syed Bukhari
University of Nottingham
MSc Computer Science
Dr. Helena Webb