DATA-DRIVEN INSIGHTS FOR PREDICTING VACCINE UPTAKE REPORT

Prepared by: **ASAM SAMWEL OLALA**

1.Poject Overview

The "Data-Driven Insights for Predicting Vaccine Uptake" project focuses on understanding the factors that influence individuals' decisions to receive the H1N1 vaccine. Using a structured data science workflow, the project examines survey data to identify key predictors of vaccination behavior and develops predictive models to support public health strategies. This study addresses the critical need to improve vaccine adoption rates by identifying barriers and enablers of vaccination. It highlights the role of healthcare accessibility, doctor recommendations, and public perceptions of vaccine effectiveness and risk. Insights gained from this analysis are intended to guide public health interventions, enhance educational outreach, and inform policy-making to promote equitable vaccine uptake.

2. Business Understanding

This project aims to explore the factors influencing the uptake of the H1N1 vaccine. By analyzing survey data, the goal is to identify key drivers of vaccination behavior and develop predictive models to assist public health officials in designing strategies to improve vaccination rates. Understanding these drivers is critical for addressing barriers and promoting public health initiatives.

The objectives of my project are;

- **Prediction**: Build machine learning models to forecast individuals' likelihood of receiving the H1N1 vaccine using demographic, socio-economic, and attitudinal data.
- **Influential Factors Analysis**: Identify key factors influencing H1N1 vaccine uptake, such as doctor recommendations, health insurance status, perceptions of vaccine effectiveness, and perceived risk of contracting H1N1.
- **Model Performance Evaluation**: Compare the effectiveness of various machine learning algorithms, including Decision Tree Classifier and Logistic Regression, in accurately predicting H1N1 vaccination behavior.

- **Insights and Recommendations**: Analyze the results of the predictive models to provide actionable insights for public health officials and policymakers aimed at improving vaccination rates.

3. Data Understanding

The dataset used in this project is H1N1_Flu_Vaccines.csv, containing data related to the H1N1 flu vaccine. The data originates from the 2009 National Flu Survey (NHFS), which aimed to capture public attitudes and behaviors concerning influenza vaccination, specifically for the H1N1 flu.

The dataset is sourced from the DrivenData competition on Kaggle (https://www.drivendata.org/competitions/66/flu-shot-learning/data/) and serves as a valuable resource for understanding the factors influencing vaccination uptake.

Key Details:
Number of Rows: 26,708
Number of Columns: 10
Target Variable:
h1n1_vaccine (Binary: 1 for vaccinated, 0 for not vaccinated)

Dataset Type: Tabular data
Data Source: 2009 National Flu Survey (NHFS)
Objective: To analyze and predict individuals' likelihood of receiving the H1N1 flu vaccine based on demographic, socio-economic, and attitudinal factors.

Column Descriptions

h1n1_concern: Level of concern about the H1N1 flu. Values: Ordinal (0: None, 1: Low, 2: Moderate, 3: High)

h1n1_knowledge: Knowledge about H1N1 flu transmission and prevention. Values: Ordinal (0: No knowledge, 1: Some knowledge, 2: High knowledge)

behavioral_antiviral_meds: Engagement in behavior related to taking antiviral medication. Values: Binary (0: No, 1: Yes)

doctor_recc_h1n1: Whether a doctor recommended the H1N1 vaccine. Values: Binary (0: No, 1: Yes)

chronic_med_condition: Whether the individual has a chronic medical condition. Values: Binary (0: No, 1: Yes) child_under_6_months: Whether there is a child under six months old in the household. Values: Binary (0: No, 1: Yes)

health_worker: Whether the individual works in healthcare. Values: Binary (0: No, 1: Yes)

opinion_h1n1_vacc_effective: Belief in the effectiveness of the H1N1 vaccine. Values: Ordinal (1: Not effective, 2: Neutral, 3: Effective)

opinion_h1n1_risk: Perceived risk of contracting H1N1 flu. Values: Ordinal (1: Low, 2: Moderate, 3: High)

h1n1_vaccine: Target variable: Whether the individual received the H1N1 vaccine. Values: Binary (0: No, 1: Yes)

4. Data Preparation

The dataset was prepared using the following steps to ensure effective analysis and modeling:

- Irrelevant Columns: Two columns, 'respondent_id' and 'seasonal_vaccine,' were removed as they did not contribute to the analysis.
- Encoding Categorical Variables: Categorical data was transformed using OneHotEncoder, which generated binary columns for each category.
- Handling Missing Values: Iterative Imputer was used to fill missing values, providing an effective solution for complex datasets.
- High-Cardinality Features: Columns with over 10 unique categories were encoded with CountEncoder, replacing category names with their frequencies.
- Splitting Data: The dataset was split into training and testing sets twice to retain a holdout set for validating model generalizability.
- Streamlined Workflow: Pipelines were implemented to manage preprocessing and modeling, reducing risks of data leakage and ensuring consistency.

5. Exploratory Data Analysis (EDA)

Key findings from EDA include:

- Doctor's Recommendation: A strong relationship was observed between receiving a doctor's recommendation and vaccination likelihood. Individuals who received recommendations had significantly higher vaccination rates. Public health campaigns should emphasize this factor.

- Health Insurance: Individuals with health insurance were more likely to receive the H1N1 vaccine (27%) compared to those without (11%). Access to healthcare coverage plays a crucial role in vaccine uptake.
- Belief in Vaccine Effectiveness: Vaccination rates increased with stronger beliefs in vaccine effectiveness. Public education campaigns should address and enhance positive perceptions about vaccine efficacy.
- Perceived Risk of H1N1: Higher perceived risk of contracting H1N1 strongly correlated with increased vaccination rates. Public health efforts should focus on raising awareness about the actual risks of H1N1 to boost vaccine uptake.

6. Modeling

Two models were developed to predict vaccine uptake: Decision Tree Classifier and Logistic Regression.

- Decision Tree Classifier:
  - AUC Score: 0.84, indicating a strong ability to distinguish between vaccinated and non-vaccinated individuals.
  - Limitations: While the AUC is high, the model exhibits low precision and F1 scores.
- Logistic Regression:
  - AUC Score: 0.84, comparable to the Decision Tree model.
  - Limitations: Like the Decision Tree, this model also has low precision and F1 scores but is not overfitting.

7. Evaluation

The models developed for predicting H1N1 vaccine uptake—specifically the Decision Tree Classifier and Logistic Regression—were evaluated using a range of performance metrics to assess their predictive accuracy and reliability. The primary metrics used for evaluation were:

- **Precision**: The proportion of true positive predictions (correctly identified vaccinated individuals) out of all positive predictions made by the model (true positives + false positives).
- **Recall**: The proportion of true positive predictions out of all actual positives (true positives + false negatives).
- **F1 Score**: The harmonic means of precision and recall, providing a balanced measure of both metrics.
- **AUC (Area Under the Curve)**: This metric reflects the model's ability to distinguish between the positive and negative classes. AUC values closer to 1 indicate better model performance.

Key Findings

- **AUC Score**: Both the Decision Tree Classifier and Logistic Regression models achieved a strong AUC score of **0.84**. This indicates that the models are generally effective at differentiating between individuals who have received the H1N1 vaccine and those who have not, with high discriminatory power.
- **Precision and F1 Score**: Despite the high AUC, both models exhibited **low precision** and **low F1 scores**. This suggests that the models struggled to make accurate predictions regarding individuals who were truly vaccinated. Low precision indicates that the models were prone to false positives (predicting someone would be vaccinated when they were not), while the low F1 score reflects an imbalance between precision and recall. This imbalance is crucial, as it suggests that while the models correctly identify some vaccinated individuals, they also misidentify many who are not vaccinated.
- **Model Robustness**: The models displayed robustness to overfitting, as evidenced by their performance across different evaluation metrics. This is an important characteristic, as it indicates that the models were generalizing well on unseen data.

Implications and Areas for Improvement

Although the models demonstrated a solid ability to predict H1N1 vaccine uptake in terms of AUC, the following areas require further refinement:

- **Advanced Algorithms**: The models could benefit from exploring more advanced machine learning algorithms, such as Random Forest, XGBoost, or neural networks. These algorithms may better capture complex patterns in the data and improve precision and recall.
- **Feature Engineering**: Improving feature engineering by incorporating additional features, interactions, or transformations may help the models better understand the key drivers of vaccination behavior. This could involve exploring more nuanced demographic and behavioral variables.
- **Addressing Data Imbalance**: The dataset may have an imbalance between vaccinated and non-vaccinated individuals, which could be influencing the model's performance. Techniques such as oversampling the minority class, under sampling the majority class, or applying synthetic data generation methods (e.g., SMOTE) could improve the model's performance in predicting the underrepresented class.
- **Hyperparameter Tuning**: Further hyperparameter optimization could help the models better fit the data, leading to improvements in both precision and recall.

By addressing these areas, we could further enhance the predictive performance of the models, making them more effective tools for guiding public health strategies, particularly in targeting populations that are less likely to receive the H1N1 vaccine.

8. Conclusions, Recommendations, and Next Steps

Conclusions

The analysis highlights critical factors influencing H1N1 vaccine uptake, including **doctor recommendations**, **health insurance access**, and **perceptions of vaccine effectiveness** and **risk.** These factors play a significant role in shaping individuals' decisions regarding vaccination.

The machine learning models developed for this project, while achieving a strong **Area Under the Curve (AUC) score of 0.84**, face challenges in **precision** and **F1 scores**, indicating the complexity of accurately predicting vaccination behavior. The low precision and F1 scores suggest that, although the models are able to identify certain patterns, they also struggle with false positives and fail to accurately predict vaccination decisions for some individuals.

Despite these limitations, the models provide valuable insights into the key drivers of vaccination decisions, informing strategies to improve public health outcomes. By understanding the factors that influence vaccine uptake, public health campaigns can be better tailored to address specific barriers and concerns, ultimately increasing vaccination rates.

Recommendations

The following are the recommendations that I have;

- **Promote Doctor Recommendations**:
  Emphasize the importance of doctor-patient communication in encouraging vaccine uptake. Training and incentivizing healthcare providers to advocate for vaccinations could significantly boost vaccination rates.
- **Address Healthcare Access Barriers**:
  Ensure vaccines are accessible to individuals regardless of health insurance status. Public health initiatives should focus on reducing financial and logistical barriers to vaccination.
- **Enhance Educational Outreach**:
  Develop targeted campaigns to address vaccine skepticism and misinformation. Educating the public about the effectiveness of vaccines and the risks associated with H1N1 can positively influence vaccination decisions.

- **Leverage Insights from Predictive Models**:
  Use the models to identify high-risk populations and tailor interventions to maximize their impact.

Next Steps

The following are the next steps that I advise be carried out;

- **Explore Additional Features**:
  Further feature engineering and integration of new data sources (e.g., recent survey data or geographic trends) can improve model accuracy and generalizability.
- **Extend Analysis to Seasonal Flu Vaccines**:
  Broaden the scope to include seasonal flu vaccination, providing a more comprehensive understanding of vaccination behaviors.
- **Experiment with Advanced Modeling Techniques**:
  Test more sophisticated machine learning algorithms (e.g., ensemble methods, neural networks) to enhance predictive performance.
- **Collaborate with Public Health Authorities**:
  Share findings with stakeholders to inform the design of real-world vaccination campaigns and policies.