Predictive Model for Tetanus Vaccination Adherence in the USA

STT811 Final Project

Diego Piraquive Gomez & Haifa Almalki

Introduction

Vaccination adherence is a critical public health concern. Due to the culture shift and growing resistance to vaccination day after day, we see a new movement asking people to stop taking vaccines[1]. Our project aims to predict whether a patient will receive the tetanus vaccine based on their demographic and healthcare data. We leverage statistical modeling and machine learning to identify key factors influencing vaccination status[2], which can help healthcare providers create better-targeted interventions and reach the high-risk demographic.

A. The Data:

Due to the challenges of obtaining original data for patients' medical history, we used synthetic data from SyntheticMass [3]. This organization provides a realistic but fictional resident of the state of Massachusetts, and this synthetic population aims to statistically mirror the real population in terms of demographics, disease burden, vaccinations, medical visits, and social determinants. It is a great opportunity for researchers in the field of applied statistics because it is free of protected health information (PHI) and personally identifiable information (PII) constraints. It is also periodically updated over time based on clinical healthcare models and epidemiological models of population health.

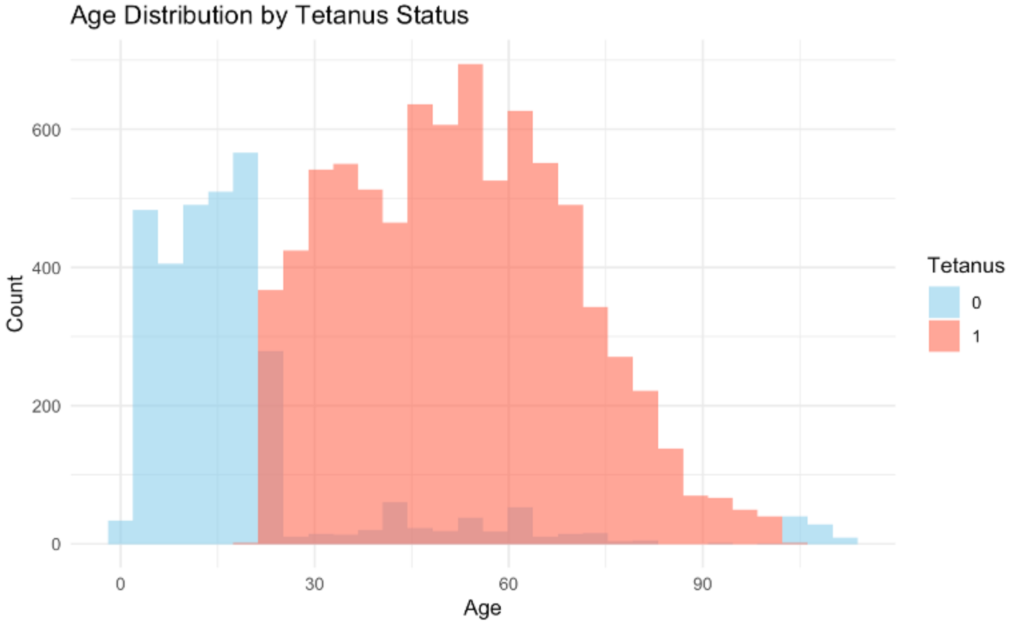
The dataset files were in Fast Healthcare Interoperability Resources (FHIR)[4] and Consolidated Clinical Document Architecture (C-CDA) format[5]. Both of these formats are standards developed by Health Level Seven International (HL7), a not-for-profit organization that works to develop standards for exchanging electronic health information between organizations. The FHIR format is widely used because it works well with mobile apps, cloud communications, and EHRs. I also support APIs, real-time data access, integration, and use in current digital health projects and apps (e.g., Apple Health). The C-CDA format is an older filing system but is widely used for standard clinical documents.

The dataset has four interrelated files that comprehensively view patient healthcare information. The Immunizations.txt file had 165,493 rows of vaccination records, including the types, dates, and other relevant info. The Encounters.txt file has 455,935 rows describing patients' visits, the date, provider profession, and visit type. While, Patients.txt has 11,363 rows that include demographic characteristics such as age, gender, ethnicity and insurance information. The Conditions file documents clinical diagnoses and health conditions identified during patient visits. SQL queries were executed to merge and manipulate the tables.

B. Exploratory Data Analysis

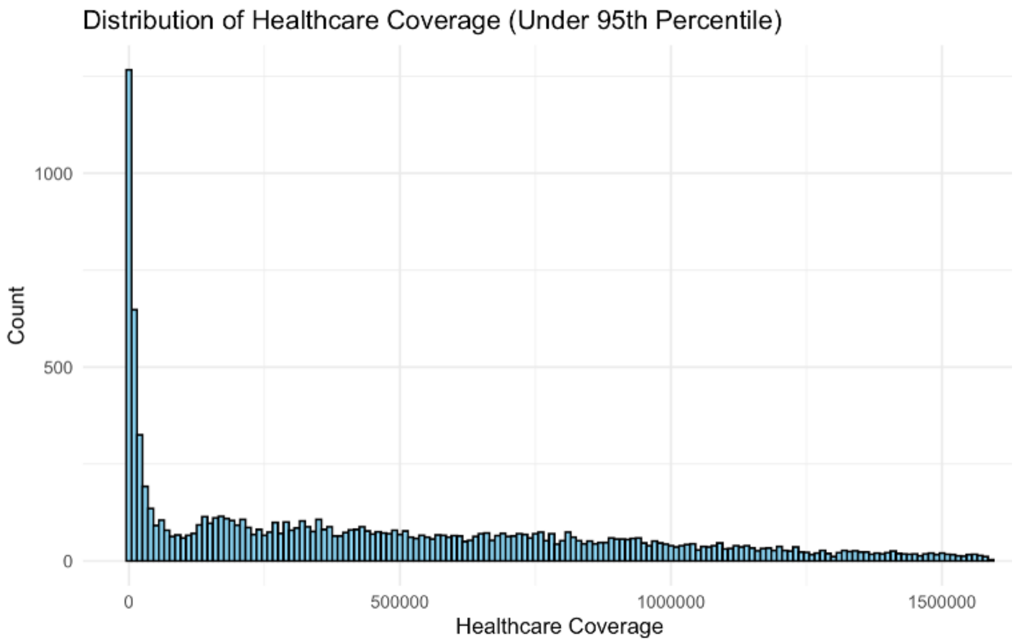
After exploring the dataset, we decided not to use with some features we considered not relevant for our predictive model, such as Social Security Number (SSN), Passport number, gender, as all the patients are women, location, since all the patients were in Massachusetts, or race, with most of the population being white. Since we are trying to build a baseline model, we wanted to keep the model as simple as possible. Now, let's examine the distribution of one of the most important features, age by tetanus status.

Figure 1. *Age Distribution by Tetanus Status*



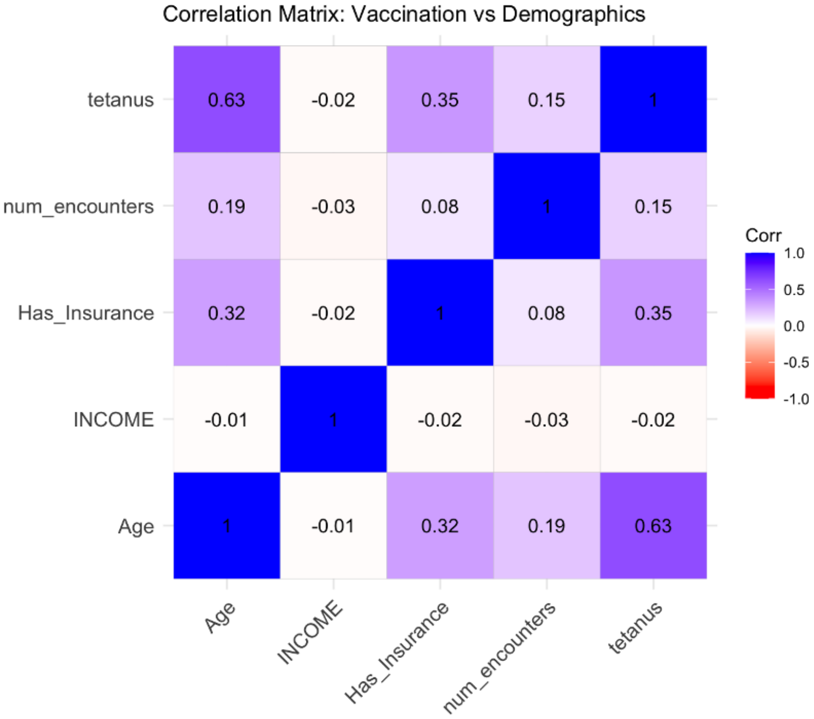
We observe that the tetanus vaccine is highly correlated with age, as most of adult patients are already vaccinated against tetanus. However, there is a small group who are not vaccinated, and they are the target population our model will try to identify to better focus the healthcare campaigns. Another relevant aspect is the healthcare coverage:

Figure 2. *Distribution of Healthcare Coverage*



It is evident that most of the population is covered by insurance. This continuous variable will help us to create a new feature indicating whether a patient has insurance, which can serve as a predictor. Finally, let’s look at our correlation matrix:

Figure 3. *Correlation Matrix: Tetanus status vs Demographics*



Based on the correlation matrix, age is the most correlated variable with the target variable 'tetanus', followed by "Has_Insurance" and "num_encounters". These variables were created through engineering, details of which will be provided detailed in the next section.

C. Preprocessing steps

During data processing, several challenges were encountered that required us to work on them prior to starting the model building. The first was embedded commas, an expected problem given our data's text format. This issue made the standard parsing methods complicated and required us to develop custom parsing functions.

For feature engineering, age was calculated using the patients' *BIRTHDATE* and *DEATHDATE* fields to capture the most accurate age at the time of vaccination or data collection. Insurance status was created as a binary variable, *Has_Insurance*, where patients with any healthcare coverage (i.e., *HEALTHCARE_COVERAGE* > 0) were assigned a value of 1. Additionally, a vaccine-specific feature was created by counting the number of tetanus doses administered per patient to better assess the immunization patterns of the patients.

Another variable, number_encounters, was created by counting the frequency of the encounters, regardless of type. We recognize that the likelihood of vaccination can be very different from a regular medical check from an emergency encounter, however, for this baseline model we are not discriminating by type of encounter.

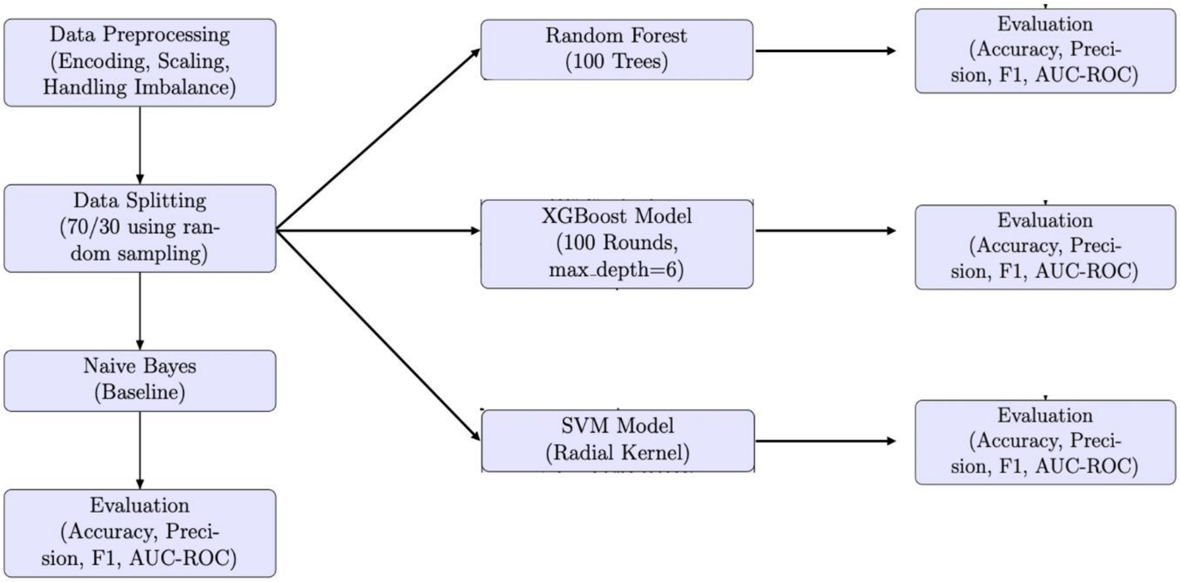
Regarding class imbalance, which is a common issue in large datasets as the number of vaccinated patients (8,198) was significantly higher than the number of unvaccinated patients (3,165), potentially affecting the performance and fairness of predictive models. For that reason, we applied SMOTE (Synthetic Minority Over-sampling Technique)[6] to oversample the minority class and avoid bias in the model.

Finally, creating the target variable was one of the most important steps. Whether a patient has been vaccinated or not with tetanus was created by first, calculating the frequency of tetanus vaccination by patiend_id in the table “Immunizations”. Then, transformed to a binary variable to just know if was vaccinated no matter the frequency, and then merged with the patient's table.

D. Model Development

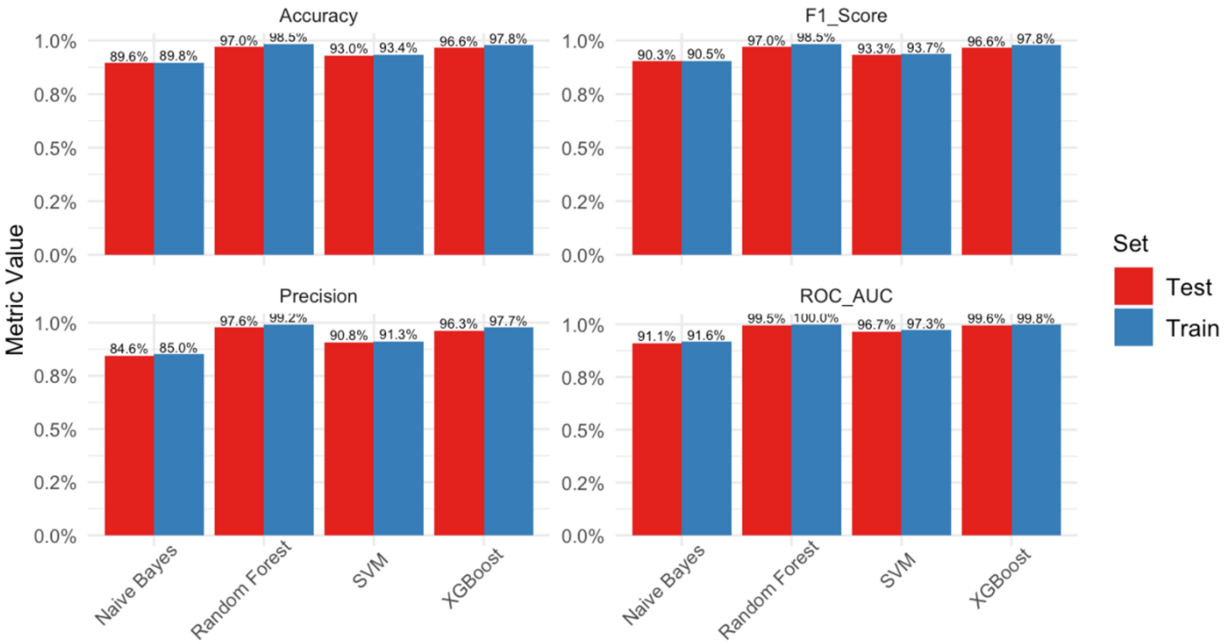
We experimented with a diverse set of classification algorithms to assess their suitability for predicting tetanus's vaccination. The models included Naïve Bayes, served as the baseline model; XGBoost, a powerful gradient-boosted tree method; Support Vector Machine (SVM) with a radial kernel; and Random Forest using 100 trees. These models were evaluated using a combination of accuracy, precision, ROC-AUC, and F1-score to capture both performance and reliability. The following pipeline (Figure 4) describes the process followed:

Figure 4. *Modeling Approach*


Data Preprocessing (Encoding, Scaling, Handling Imbalance)
Data Splitting (70/30 using random sampling)
Naive Bayes (Baseline)
Evaluation (Accuracy, Precision, F1, AUC-ROC)
Random Forest (100 Trees)
XGBoost Model (100 Rounds, max_depth=6)
SVM Model (Radial Kernel)
Evaluation (Accuracy, Precision, F1, AUC-ROC)
Evaluation (Accuracy, Precision, F1, AUC-ROC)
Evaluation (Accuracy, Precision, F1, AUC-ROC)

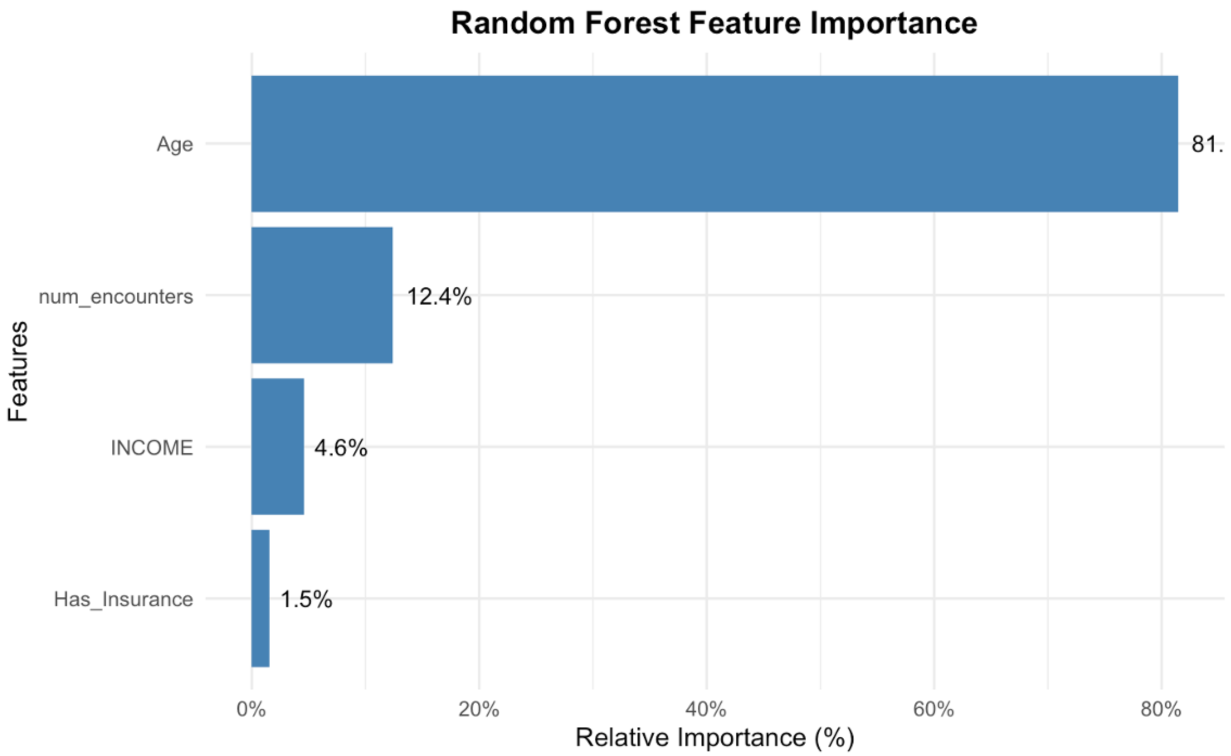
The following figure shows the model performance results:

Figure 5. *Model Performance Results*



In terms of accuracy, our baseline model performed at 89.6%. However, the best performers, Random Forest and XGBoost, achieved 97% and 96.6% accuracy in the test set, respectively. Similarly, these models showed superior precision, indicating their effectiveness at minimizing false positives. This is also reflected in their F1 scores, confirming that all models balance well between false positives and false negatives. Regarding ROC AUC, it appears all models discriminate well between classes. The consistent performance across both training and test sets suggests minimal overfitting, as models perform well on both trained and unseen data.

Figure 6. *Random Forest Feature Importance*



As expected, age was the most significant factor identified by the Random Forest model, followed by num_encounters and income. It is important to note that this finding is specific to the tetanus vaccine, which is highly correlated with age, particularly within the population of Massachusetts.

Conclusion and Future Work

In conclusion, this study demonstrated that machine learning models can effectively predict tetanus vaccination status using demographic, clinical, and insurance-related features. After addressing data challenges such as embedded commas, class imbalance, and missing values, feature engineering, and exploratory data analysis revealed strong correlations between age vaccination status and high insurance coverage among patients. While the model was limited by geographic bias (Massachusetts-only data) and a binary classification approach that did not capture partial vaccination, future work could expand the model to include other vaccines and incorporate temporal trends. Furthermore, in terms of future technical work, hyperparameter tuning could be implemented to enhance results when analyzing other vaccines. In this project, with the tetanus vaccine, it was not implemented due to the strong performance of our baseline model.

Code file :STT811_PROJ/STT_Project.Rmd at main · diegopiraquive/STT811_PROJ

Ethical Statement

ChatGPT-4 , DeepSeek and Grammarly were used to refine the writing improve the idea and making the statement shorter. We used ChatGPT-4 to structure our report and help us cover all the required steps.

Reference

1. Shen, S. (Cindy) & Dubey, V. Addressing vaccine hesitancy. *Can Fam Physician* **65**, 175–181 (2019).
2. AlShurman, B. A., Khan, A. F., Mac, C., Majeed, M. & Butt, Z. A. What Demographic, Social, and Contextual Factors Influence the Intention to Use COVID-19 Vaccines: A Scoping Review. *Int J Environ Res Public Health* **18**, 9342 (2021).
3. About | Synthea. https://synthea.mitre.org/about.
4. FHIR® - Fast Healthcare Interoperability Resources® | eCQI Resource Center. https://ecqi.healthit.gov/fhir?qt-tabs_fhir=about.

5. Consolidated Clinical Document Architecture (C-CDA). https://site.healthit.gov/c-cda.
6. SWASTIK. SMOTE for Imbalanced Classification with Python. *Analytics Vidhya* https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/ (2020).