National College of Ireland

Predictions of Changes in Child Immunization Rates Using an Automated Approach: USA

MSc Research Project
Data Analytics

Niall Mannion
Student ID: x17166985

School of Computing
National College of Ireland

Supervisor: Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …Niall Mannion……

Student ID: …x17166985……..……

Programme: …MSc Data Analytics…… Year: …2020……...

Module: …MSc Research Project…….……

Supervisor: …Catherine Mulwa…….……

Submission Due Date: …28/09/2020……..……..……..

Project Title: … Sentiment Analysis and Predictions of Changes in Child VRs Using an Automated Approach: USA……

Word Count: …… Page Count……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictions of Change in Child Immunization Rates Using an Automated Approach: USA

Niall Mannion
X17166985

Abstract

Immunizations help to save millions of lives each year from infectious diseases. Declining rates of immunizations caused by concerns about their safety and sociodemographic/socioeconomic factors have resulted in a rise in deaths from infectious disease. Past studies in the area of immunization hesitancy have focused on generating advanced statistical and machine learning techniques that non-coders may not be able to use. This research designed a platform using R Markdown that analysed immunization related content from Twitter and Google Trends, generated data mining models using patient survey data to predict whether or not a child will receive immunizations, and predicted trends in immunization coverage with time series forecasts. The work generated a visual dashboard of the online content, and the SuperLearner data mining model was the most accurate model in predicting child's immunization status, with 76% accuracy, sensitivity of 30.01% and specificity of 80%. Timeseries forecasts had a mean absolute error of 6.83. The simple automated platform could be used by healthcare workers and public health officials to model trends in immunization coverage and anticipate changes in immunization rates to prevent deaths from infectious disease.

1 Introduction

This introduction details the previous work in the area of using data mining to analyse and predict immunization rates and immunization preventable diseases, explains why the project aimed to address this issue and why the study and its resultant platform are of value to healthcare professionals. The study aimed to provide healthcare workers with a user-friendly platform that could be used to model changes in immunization rates, which could help predict infectious disease outbreaks and reduce mortality from infectious disease.

According to the World Health Organisation 2-3 million lives are saved each year worldwide due to immunizations, but 1.5 million extra lives could be saved every year by improving immunization uptake (Okwo-Bele, J.M., 2019). Suboptimal immunization levels are driven by negative sentiments to immunizations online (Tustin et al., 2018), discredited studies such as Andrew Wakefield's study which found a link between immunizations and autism (Kolodziejski, 2014) and other factors such as poverty and lack of access education and healthcare (Larson et al. (2016).

1.1 Project Background and Motivation

The COVID-19 pandemic has shown the threat that infectious disease has on human life. The outbreak has caused hundreds of thousands of deaths and the shutting down of economies, but there may be hope for the eradication of the disease through immunizations. Researchers at institutions such as Oxford University have been developing an immunization to coronavirus, which have shown efficacy in animal trials (Palca; 2020). Immunizations may help eradicate COVID-19 but even if an immunization is developed concerns about the safety of the immunization in the general population may prevent patients from receiving the immunization and prolong the pandemic. Herd immunity is when a large percentage of the population is immunized against a disease which prevents the spread of the disease even to a person who is not immunized (National Geographic; 2016). Small reductions in immunization rates can prevent herd immunity and cause infectious disease outbreaks and mass.

The aim of this research was to provide healthcare workers with a platform called the Immunization Analysis Platform (IAP) that can be used to analyse immunization related online information, predict if a child is going to be immunized and forecast future rates of immunization using historical data. The platform aims to provide healthcare workers with a tool that is user-friendly and automatically generates interactive dashboards to provide insight into anti-immunization attitudes and into the socioeconomic factors that contribute to a parent's decision to not immunize their child. The tool could help doctors understand the problem of low immunization uptake and intervene in cases where the child is likely to default from immunization programs. For example, if a Coronavirus immunization is developed the tool could be used to extract, analyze and visualise online content related to the immunization, the predictive models could be used to predict the likelihood of individuals getting immunized based on whether or not they had been immunized against other diseases, and time series forecasting could be used to predict the uptake of the immunization in different regions based on how many people have been immunized already.

Infectious diseases that can be prevented by immunizations occur most often in areas with low immunization rates. In 2019 a measles outbreak in New York hit an area where 89% of the population was either unimmunized or had an unknown immunization status (CDC, 2019(a)). The aim of this project is to allow healthcare workers to identify these hotspots of low immunization coverage to predict future disease outbreaks and prevent needless deaths.

1.2 Project Requirement Specification

1.2.1 Research Question

The research question for this study addresses the need to improve immunization uptake in countries such as the United States. The COVID-19 pandemic has shown that in todays' globalized world infectious diseases can spread and lead to a large mortality rate very quickly. Future research may lead to a COVID-19 immunization which could eliminate the disease. However, the uptake of this immunization may be reduced because of negative attitudes towards immunizations and social factors such as poverty and poor healthcare access. The study aimed to design an automated data mining platform because immunization related data is often complex, large in volume and unstructured. This makes the task of understanding declining immunization rates difficult, especially for those without programming experience. The target stakeholders for this research were healthcare workers. R Studio, Flexdashboard,

Shiny, SQL Server Management Studio (SSMS) and SQL Server Integration Services (SSIS) were used to build the IAP.

Research Question: "To what extent can data mining modelling (such as SVM, KNN and Random Forest) enhance prediction of the likelihood of a child being immunized in the USA to reduce mortality rate"

Sub-RQ: "To what extent can Twitter, sentiment analysis using TidyText and Google Search Volume data; be used to understand attitudes towards immunization and can time series forecasting improve predict immunization rates?"

1.3 Research Objectives and Contributions

The research objectives for the project are shown in table 1.

Table 1: Project Research objectives

Objectives	Description	Evaluation Metrics
1	Generate an IAP dataset that contains tweets and Google search volume data, an immunization survey of parents and a timeseries of past immunization rates in the US	
2	Perform preprocessing, exploratory analysis and feature selection on the IAP data	
3	Perform sentiment analysis and text analysis on the tweets using TidyText and visualize the Google trends data	
4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 4.10	Implement machine learning algorithms to predict childhood immunization status Build Naïve Bayes models and evaluate results Build C5.0 models and evaluate results Build Random Forest models and evaluate results Build Support Vector Machine models and evaluate results Build C-Forest models and evaluate results Build KNN models and evaluate results Build Neural Network models and evaluate results Build Bagging models and evaluate results Build Boosting models and evaluate results Build SuperLearner models and evaluate results	Specificity, sensitivity, accuracy
5	Compare the performance of the machine learning models together and with previous work	
6	Implement timeseries forecasts with forecastML	Mean absolute error (MAE), Root mean square error (RMSE)

Contributions: The major contribution of this research was to provide healthcare workers with a user-friendly platform that could be used to analyse data related to immunizations without needing programming experience. The IAP can be used to analyse online content related to immunizations, predict the likelihood of a child being immunized and predict future rates of immunization based on historical rates. Salmon et al. (2015) note that predicting which geographical locations will be hotspots for low immunization rates, analyzing attitudes to immunization to understand anti-immunization sentiment and predicting which children are likely to default from immunization programs could help improve immunization uptake and reduce the likelihood of disease outbreaks. Globally 17% of deaths in children under 5 are caused by immunization preventable diseases, which highlights the importance of understanding and improving immunization rates to protect children's lives

The rest of the study is structured as followings: Section 2 contains a summary of the recent work in the area of epidemic and immunization rate prediction; Section 3 discusses the methodology of the project; Section 4 presents the implementation, evaluation and results stages; Section 5 discusses the results and implications of these findings; finally, Section 6 contains the conclusion and future work section.

2 Literature Review on Analysing Immunization Related Data and Predicting Immunization Rates (2010 –2020)

This section discusses the previous literature related to the topic of immunization prediction. In section 2.2 the use of data from online sources such as Twitter and Google is discussed. Section 2.3 describes the literature that was used to select the other features in this project, such as socioeconomic and demographic factors. Section 2.4 discusses how the machine learning techniques and evaluation criteria were selected for this study, and section 5 discusses the use of automation with healthcare data.

2.1 A Review of Social Media Relationship with Disease Outbreaks and Vaccines

One of the objectives of this study was to give healthcare workers a platform that extracts and analyses online content to gain an insight into how online users feel about immunizations. Several previous studies have used Twitter to predict the spread of infectious disease and analyse attitudes towards immunizations. Hayate et al. (2016) scraped Twitter posts that contained words such as “fever” and “headache” to forecast (using linear forecasting) the spread of influenza. They geolocated tweets and mapped outbreaks based on these symptoms. The strength of the study was that it had a very large sample size of 7.7 million Twitter posts over the course of four years, and the results were strong, with a 0.91 correlation between the frequency of tweets related to symptoms and flu rates 1 week in the future. They also found a 0.77 correlation when predicting rates three weeks in the future. One of the issues with this study is the lack of real-time streaming of tweets, which limits the user’s ability to predict rates in real time.

Achrekar et al. (2011) addressed this issue by providing a framework that streams Twitter posts continuously. They built auto-regression models that predicted flu rates based on the frequency of posts such as “I have flu” or “down with swine flu”. Their models had a correlation of 0.89 between flu rates in an area and the frequency of flu related tweets. These studies show the

value of social media in analysing the spread of infectious disease, but they do have limitations. The frequency of words such as “flu” or “fever” may only begin to increase when an outbreak has begun. As has been seen in outbreaks like the measles outbreak in New York which led to 649 cases (Zucker et al. 2020). Low adoption of immunizations can be a more useful predictor of disease than data that focuses on those already infected.

Other researchers used low immunization rates as a surveillance system for epidemics. Dunna et al. (2017) analysed the role of the media in influencing immunization rates. The strength of the study was that they used a very large sample size of 250,418 tweets combined with socioeconomic factors. They used regression to predict immunization rates and mapped people's exposure to articles related to the HPV vaccine. The exposures accounted for 17% and 12% of the variance immunization rates in men and women respectively.

Another issue with this previous work in the field is that it fails to take the emotions of users into account. A post such as "immunizations are dangerous" and "immunizations are very effective" both mention immunizations but express different emotions. Salathe et al. (2011) performed sentiment analysis on immunization related tweets and analysed the correlation between immunization rates in an area and sentiment towards immunizations in that area. The strength of the study was that they used a large sample size of 470,000 and took the sentiment of the posts into account. They found that there was a 0.78 ($p=0.02$) correlation between sentiments at a regional level and immunization rates and 0.52 ($p=0.005$) at the state level.

Another study by Tomeny et al. (2017) used sentiment analysis and found that tweets from states such as California were more likely to be negative towards immunizations. A strength of this study is that they incorporate other factors such as socioeconomic factors into their analysis of immunization rates rather than just focusing on online content. Based on census data they found that poorer individuals, men and younger people were more likely to have positive attitudes towards immunizations. A limitation of these studies however is that they focus on immunization related tweets, which doesn’t take into account attitudes outside of Twitter. Blank et al. (2016) found that Twitter users are generally younger and have a higher income than the national average, which means using just Twitter data creates skewed samples. Mavragani et al. (2018) analysed interest in immunizations and predict immunization rates using Google Search data. They found that there was a negative correlation of -0.76 ($p < 0.01$) between interest in the phrase “anti-vaccine” and immunization rates in different regions.

One issue with this previous work is that they focus on online posts in the English language. Approximately 20% of the US consists of immigrants whose main language is not English (Burton et al., 2018). Immigrants and non-English speakers are often excluded from healthcare studies because these studies are often implemented by English speaking researchers (Garrett et al., 2010). In studies related to immunization, this could lead to low immunization rates in migrant communities being ignored by predictive machine learning models. Findings from groups such as Lu et al. (2015) that races such as black and hispanic communities have lower immunization rates than white communities. This study aimed to overcome this limitation by extracting content from Spanish speaking users and translating this into English in R.

The search terms for the study were chosen based on previous work. Kang et al. (2017) used a semantic network to visualise the most frequent words in immunization related articles retweeted by Twitter users and the sentiment of these articles. They found that articles that had a negative sentiment tended to contain words such as “thimerosal” and “mainstream media”.

Articles with a positive sentiment were more likely to contain words such as “autism”. This provided an insight into the most popular keywords used in popular immunization related articles (both positive and negative), and a number of these keywords were incorporated into this study. Based on this previous work this study aimed to provide healthcare workers with a framework that could provide a continuous flow of online content that could be extracted and analysed automatically. The framework analyses immunization related data by translating tweets into English, combining analysis of online content with the analysis of other factors such as socioeconomic factors and Google Trend data.

2.2 Data and Feature Selection

Many predictive factors were used in this research to understand trends in immunization and to predict the likelihood of a child being immunized. De Figueiredo et al. (2016) found that factors such as ethnicity, income levels and education level influenced whether people got immunized. They predicted worldwide rates of immunization using regression models. The strength of the study was that it took a large sample size, looking at immunization rates across 190 countries over 130 years. The large number of countries studied allowed regional differences to be considered, showing the complex nature of the problem of immunization uptake. Low income was associated with reduced immunization uptake in some regions, in regions such as the Mediterranean lower government spending led to lower immunization rates, whilst in Africa finishing primary school was associated with higher immunization rates.

Another study by Hu et al. (2014) looked at the factors that effected immunization rates in China. One strength of the study was that they considered many factors which effected immunization rates. This project aimed to follow a similar approach by considering many factors in order to build the best predictive models. Hu et al. (2014) used regression and Chi-Squared tests to analyse the relationship between various factors with immunization rates and found maternal age, average income level of the area and education level all influenced immunization status. One issue with this and De Figueiredo et al.’s (2016) work is that the models that were built were simple statistical models that did not have predictive power.

A study by Spencer et al. (2014) measured if there was a correlation between immunization rates and poverty. High levels of poverty were associated with lower uptake of the HPV vaccine. A study by Larson et al. (2016) used surveys across 67 countries to analyze the relationship between factors such as poverty and education with immunization rates. Like Figueiredo et al. (2016) the study looked at many regions and had a large sample size of 65,819 respondents, which improves the confidence in the findings. They used logistic hierarchical techniques and found that higher educational attainment and improved healthcare access led to an increase in anti-immunization sentiment. Other studies such as Smith et al. (2010) also found that higher education levels led to an increased chance of that person opposing immunizations. This shows the complex nature of the problem, with different studies showing contradictory findings that may be dependent on other factors such as regional differences.

One issue with work such as De Figueiredo et al. (2016) is the lack of explanation for why people are hesitant about immunizations. A study by Polonijo et al. (2013) addressed this problem by using logistic regression models on National Immunization Survey (NIS) datasets related to HPV immunizations. They used factors such as if the individual had heard of the HPV immunization, if a healthcare worker encouraged them to immunize against HPV and included this with other factors such as the persons’ income and race. They found that many

people didn't get the immunization because they had not been advised by their doctor about the vaccine, and found that those from racial minorities and with a lower income were less likely to have heard of the vaccine. Another study by Freed et al. (2010) found that 11.5% of women had refused at least one recommended immunization for their child, and many expressed concerns about the adverse effects of immunizations. This shows the important role that doctors play in reaching out to vulnerable populations and advising them to immunize. This project aims to help doctors with this issue to identify these communities with low immunization rates and encourage them to get immunized. Taksdal et al. (2013) found that local doctors encouraging their patients to immunize their child increased the chances of the child getting a flu shot. This project aimed to give doctors a tool that could be used by doctors to mine insight from social media and Google Search trends to understand why parents are hesitant to immunize their child.

One of the purposes of this study was to provide an understanding of the factors that influenced immunization status in children. A study by Canavan et al. (2014) used logistic regression to analyse the relationship between socioeconomic factors and immunization rates in children from East Africa. Their results suggested that those delivered in a public or private hospital were more likely to get immunized than those delivered at home, suggesting that access to healthcare may influence immunization rates. This project includes multiple factors related to healthcare, namely insurance status, healthcare provider type and number of providers as candidate predictors for immunization status. One issue with the study by Canavan et al. (2014) is that they only use a logistic regression model, which may not notice trends in the data that more complex models do. This project will also use algorithms such as deep learners to address this issue.

2.3 Literature Review on Model and Evaluation Metric Selection

The machine learning algorithms and criteria used to assess their performance were selected based on previous work in healthcare research. Many of the previous studies into the factors effecting immunization rates have been simple nature and the lack of predictive power machine learning models. Pavlopoulou (2013) and Crouch and Dickes (2015) measured the statistical relationship between demographic and socioeconomic features with immunization rates. These models are relatively simple and can often be less accurate than more complex techniques such as deep learners. Pavlopodou et al. (2013) also used a relatively small sample size of 731 patients. Larger sample sizes often produce more accurate results.

Chandir et al. (2018) addressed this limitation by using a much larger sample size of 47,554 rows. They also compared four machine learning algorithms to test which best predicted immunization status in children. One strength of the study over some previous studies is that they built predictive models. This meant that their study had a practical purpose in that it allowed doctors to predict if a child was immunized rather than just researching the factors that influence immunization status. The study was the most similar to this project and therefore the results of this project were compared with the work by Chandir et al. (2018). They used Random Forest, SVM, Decision Trees and C-Forest and Random Forest performed best with an accuracy of 79.1%, sensitivity of 94.9% and specificity of 54.9%. This project aimed to use a large dataset and use techniques that are suited to handling larger datasets such as deep learners. This project also compared many different machine learning techniques just like Chandir et al. (2018). It built on this work by offering doctors an automated tool, implementing more algorithms including deep learners such as Neural Networks and includes datasets from

other sources such as online content to analyse the underlying attitudes that lead to anti-immunization sentiments.

The algorithms used in this study were chosen based on a number of previous studies. A study by Huang et al. (2015) identified deep learners as the best algorithms for analyzing healthcare datasets in public health research. Deep learners are suited to deal with large, complex datasets with many variables and generally produce highly accurate models with this data. Research by Shi et al. (2012) found that neural network models were more accurate in predicting the mortality of those recovering from cancer procedures than logistic regression models and Padmavathi et al. (2011) found that neural networks performed better than logistic regression and multilayer perceptron algorithms when predicting the incidence of breast cancer. Specificity and sensitivity are commonly used in these healthcare studies and will therefore also be used in this study. Research by Signorini et al. (2011) predicted the spread of the H1N1 pandemic using Twitter posts. They found posts containing keywords such as “vaccine” and “influenza” and used the locations of these tweets to predict where the virus was spreading. Their research used SVM models to predict the spread of the pandemic with an error of 0.28%.

2.4 Automation in Healthcare

The purpose of this study was to provide healthcare workers with an automated system that could extract, clean, analyse, forecast and visualise data and predict a child's immunization status without the need for programming experience. One issue with previous studies in this area is the lack of practical application of the work. Bhattacharya et al. (2010) automated the task of using data mining models on hospital patient logs and managed to obtain detailed patient information and reduced time spent on the manual creating of medical logs. Anguera et al. (2016) used data mining to assess time series data consisting of patient brain scans and this approach better highly accurate models for predicting epilepsy. Another study by Xie et al. (2018) used an automated system to identify scalpers (those who bought services to sell at a higher price) in the healthcare industry. They used clustering to obtain an accuracy of up to 93% in identifying scalpers. These studies are well designed because they focused providing end users such as doctors and nurses with simple, automated tools to improve their work and reduce time consumption. They also allow non-programmers to use complex data mining techniques. This is why an automated approach was used for this study.

2.5 Identified Gaps

This section discussed the previous work related to the chosen topic of this research. This study aimed to add to these studies by demonstrating the use of automated systems in healthcare analytics. Unlike previous work this study combined multiple sources of data such as online data, socioeconomic and demographic data and historical time series data to help doctors understand the issue of immunization coverage, identify those likely to default from immunization programs and predict future trends in immunization rates. The following chapter discusses the methodology that was used in this study and explains the design decisions that were made for the project.

3 IAP Methodology Approach and Design Used

3.1 Introduction

In this section the Knowledge Discovery Database (KDD) research methodology (Fayyad, 1994) was followed. The motivation behind the project was to create an automated platform

that healthcare workers could use to analyse trends in immunization coverage, to understand and monitor attitudes towards immunizations and to identify the factors that influence immunization status and use these to predict whether or not a child would get immunized. A three tier architecture was used in order to provide a persistent layer for the social media data, which could only be scraped on a weekly basis.

3.2 IAP Methodology Approach Used

The research approach used for the project is shown in Figure 1.

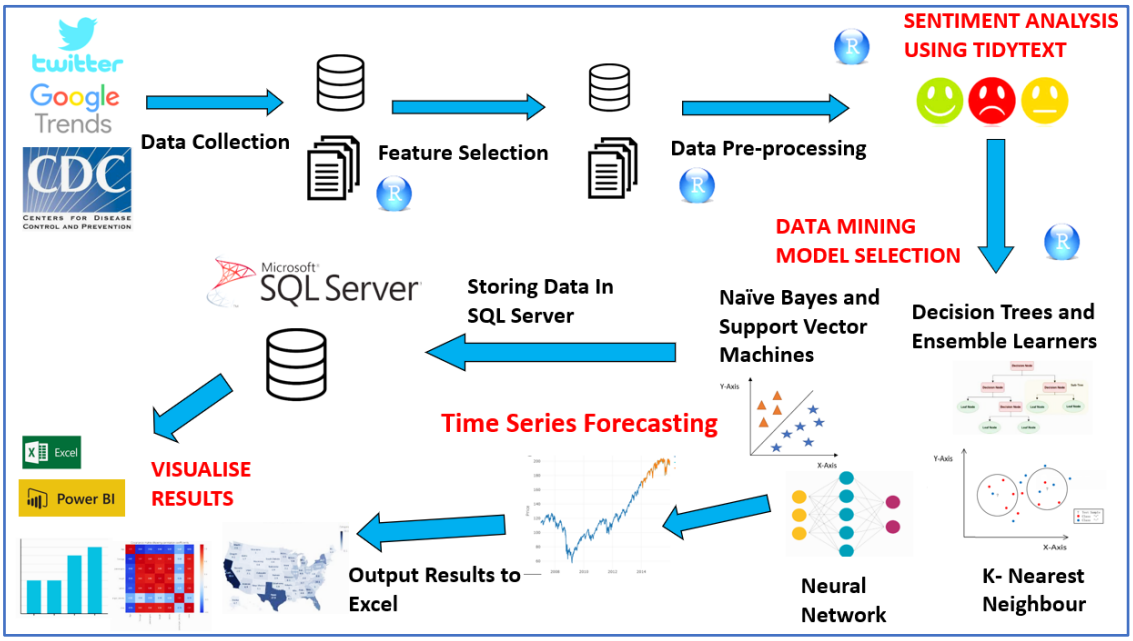

Figure 1: Immunization methodology approach

3.2.1 Data Selection

The source of the NIS data for the years 2017 and 2018 was the CDC (CDC; 2019(c)). The 'rtweet' package was used to extract immunization related Twitter posts from the previous week from the US. 'gtrendsR' was used to extract Google Search data related to immunizations. This data contained Google search trends for keywords related to "vaccines" (because this was a common search term than "immunization") over the period that Google has operated (around 16 years) and across each US state. The timeseries data of child immunization rates in the US (1995-2017) came from the CDC (2018). These datasets were saved as a csv file and then persisted on SSMS with the SSIS software.

3.2.2 Exploratory Analysis

The NIS dataset was analysed using exploratory analysis to understand the underlying nature of the dataset. T-Tests and Chi-squared tests measured the relationship between numerical and categorical variables with the outcome variable, immunization status. Histograms and boxplots were used to visualize distribution and outliers respectively. Plots were generated to compare the immunization status of children based on different factors such as parent's income and family size. Plots were also created to view the ratio of unimmunized to immunized children.

3.2.3 Data Pre-processing

The datasets were processed using several steps. The Twitter posts were cleaned. The Twitter dataset was changed to the appropriate classes (text posts converted to string etc.) and the text was transformed into a document term matrix. Rows with nulls values were taken out and Spanish language text was translated into English. The NIS survey datasets from the two years were combined, rows with null values removed and columns were changed to the correct class. The time series dataset had outliers that were normalised using the 'tsoutliers' R package.

The class imbalance in target variable (whether or not the child was immunized) was corrected in R using the "ROSE" package. Four techniques were used: oversampling- duplicates rows from the unimmunized chidlren; undersampling- removes some of the samples of immunized children; synthetic- artificial samples of the unimmunized class are created; mixed sampling- uses a mixture of these methods to increase the number of unimmunized samples.

3.2.4 Feature Selection

Random Forest, Chi Squared tests and the Boruta algorithm were used to choose the features for the machine learning algorithms. Previous work was used to identify features that would be strong predictors of immunization status. These features mainly consisted of socioeconomic factors such as income level, level of education and type of healthcare (public/private), as well as other factors such as state of residence, household size and mother's age group. Visuals were also generated using the 'ggplot2' package in R to help visualize the relationship between immunization status and the different features selected.

3.2.5 Data Mining

Ten data mining techniques were used in this study to predict the immunization status of children: SVM, KNN, SuperLearners, Random Forest, C-Forest, Neural Networks, Boosting, Decision Trees and Bagging and Naïve Bayes. The criteria used to evaluate the performance of the models were specificity, sensitivity and accuracy.

3.2.6 Data Interpretation and Evaluation

Graphs of the online data were generated that showed findings such as the most frequently used words in the Twitter data, the sentiment of the most commonly occurring words, the overall sentiment of all of the tweets (positive or negative) and Google Search volumes for each state. The performance of each machine learning model used to predict immunization status was plotted, and forecasts for outliers and predicted values were generated for the time series data. Each of these plots were generated using 'ggplot2' in R.

3.3 Design Specification Process Flow

The project design process (shown in figure 2) consisted of a layer for visualizing the data, a second tier consisting of the logic tier where machine learning models were built to analyse the sentiment of the text data, predict immunization status and forecast trends in immunization coverage. This tier also included steps for processing and transforming the data, exploratory analysis and feature selection and the evaluation of each model. The final tier was a persistent tier in SSMS which stored the project data.

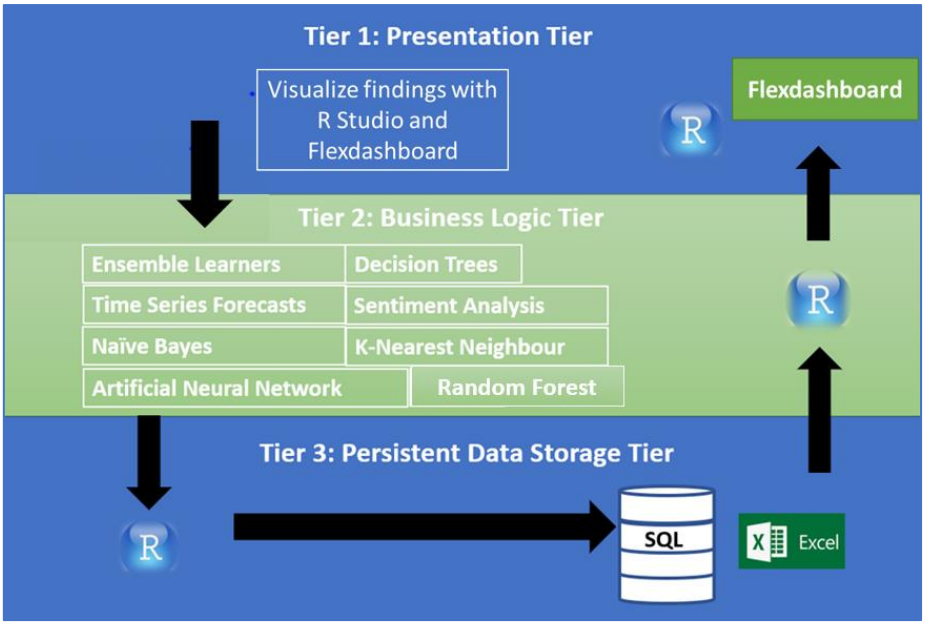

Figure 2: Architecture Design of Immunization Prediction Platform

3.3 Conclusion

This study followed the KDD methodology approach. Datasets from online sources, survey data related to immunization status of children and historical immunization rates in different US regions were loaded into R Studio, preprocessed and transformed. The online data was analysed using visualizations, token unnesting and sentiment analysis. Exploratory analysis was performed on the NIS data, and machine learning algorithms were built using R to predict the immunization status of a child using socioeconomic and demographic features. Predictive forecasting was applied to the data containing immunization rates for three states in the US and generate plots of these forecasts. An interactive dashboard was generated using R Studio, Shiny and Flexdashboard to visualize these findings.

4 Implement, Evaluation and Results of Immunization Analysis Models

This section describes the implementation, evaluation and results chapters for the study. The design decisions that were made for the project, the technologies, performance metrics and steps such as exploratory analysis and feature selection are presented in this chapter.

4.1 Extraction and Description of Datasets

There were three main stages of the project. The first stage was the exploration and analysis of immunization related online content from Twitter and Google Trends. Tweets from the US that contained the keywords “immunization”, “immunized”, “vaccinated”, “vaccination” and “vaccine” were extracted using TwitterR in R. Tweets from the previous week were scraped. ‘GtrendsR’ was used in R to scrape search trends for keywords such as “vaccine” and “immunization” over the last 16 years and the interest in vaccines for each state in the US in the last 12 months..

Two survey datasets for child immunization status were obtained from the CDC (2019(c)). They contained a questionnaire for parents from the US for the years 2017 and 2018, asking

whether they had immunized their child against the most common infectious diseases. There were also a number of other factors that were documented, such as the size of the household, the child's race, the income of the parent's and the state they live in. A description of the factors that were included in the project is shown in table 2 in the configuration manual. A column was added to the table to show if the factor had a statistically significant effect on immunization status. The predicted variable in the study was the child's immunization status. There were a large number of columns for different immunizations in the dataset, and the column containing the '431-313' values was the chosen outcome variable. This column contained whether or not the child had received immunizations against the most common childhood infectious diseases, such as measles, mumps and diphtheria.

The source of the time series data was the CDC (2018). This contained immunization rates for different regions of the US between 1995-2017. The dataset rows containing the immunization rates for Texas and two of it's neighboring states (Oklahoma and Arkansas) were extracted from the dataset because this study focused on Texas and nearby regions.

4.2 Data Preprocessing

The tweets were prepared for text analysis by removing contractions (e.g. converting "won't" to "will not"), punctuation such as hashtags, commas and question marks were taken out, upper case letters were changed to lower case, each word was unnested (sentences broken up a list of each word) and Spanish text was converted to English using the translateR function. The NIS data was preprocessed by removing rows containing missing values or entries such as 'N/A', removing rows where adequate provider data wasn't provided (i.e. the doctor hadn't given enough data for the patient), and features that were thought to be predictive of immunization status based on previous work were retained in the dataset. The data was imbalanced due to the relatively high number of immunized children, and this was corrected using the 'ROSE' package which oversampled the children that weren't immunized. The CDC data used for forecasting did not need to be changed from it's raw form.

4.3 Exploratory Analysis

The 'TidyText' R package was used to analyse the immunization related tweets. The most frequently occurring words in the tweets and the emotions expressed by the users and the emotions of the most commonly occurring words in the tweets were found using token unnesting and sentiment analysis, and plots were generated using the 'ggplot2' package to present visuals of these analyses. The Google search data was also visualized using 'ggplot2', with the Google search interest over time for each keyword and the interest by region visualized using a 'ggmap' function with the US state map.

Results and Evaluation:

The tweets were from the 3rd to the 10th of August. The most common words in the Twitter data are shown in Figure 2. 'covid' is the most commonly occurring topic, appearing over 500 times. This suggests that there was large interest in the coronavirus vaccine, as evidenced by the high frequency of words such as 'coronavirus' and 'virus', as well as the link between politics and vaccines with 'Trump' being another commonly occurring word. 'russia' and 'putin' also occur frequently, which may be due to the recent news that Russia have begun development on a coronavirus vaccine (Foy, 2020).

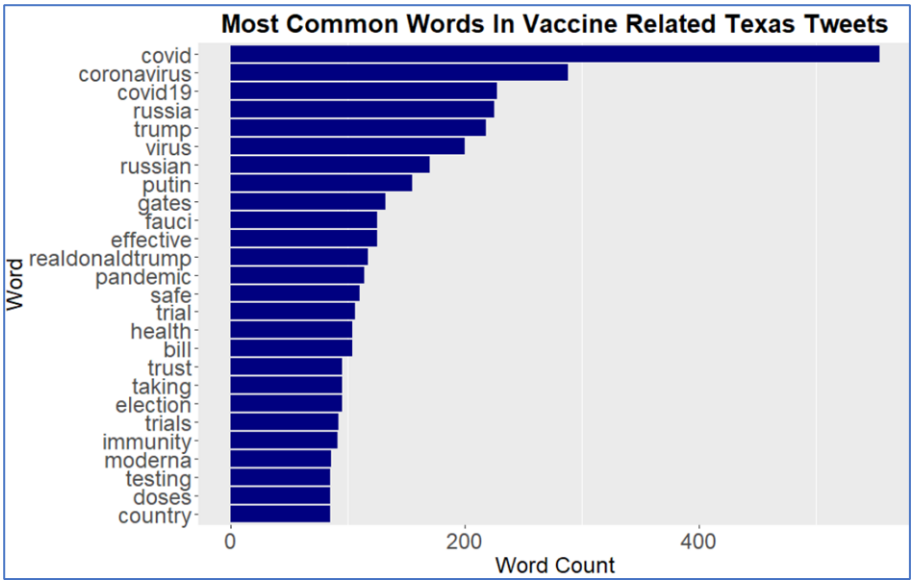

Figure 3*:* Most frequently used words in vaccine related tweets from the US

The most common emotions expressed by the Twitter posts are shown in figure 133 in the configuration manual. Trust was the most common sentiment amongst users, with fear being second.

A line graph is shown in figure 4 of Google search volume for keywords related to vaccines (which were more commonly searched than immunizations). The graph shows there were spikes in interest in 2009 and 2020, which coincides with the rise of the swine flu and COVID-19 pandemic. This suggests that users turn to the internet during pandemics to search for information related to vaccines. This may help to explain why anti-immunization sentiment is so common. Users may be getting incorrect information related to immunizations from online sources, and this may be prevented by organizations such as the CDC or WHO posting advertisements for information on immunizations that tell people of the benefits of immunizations.

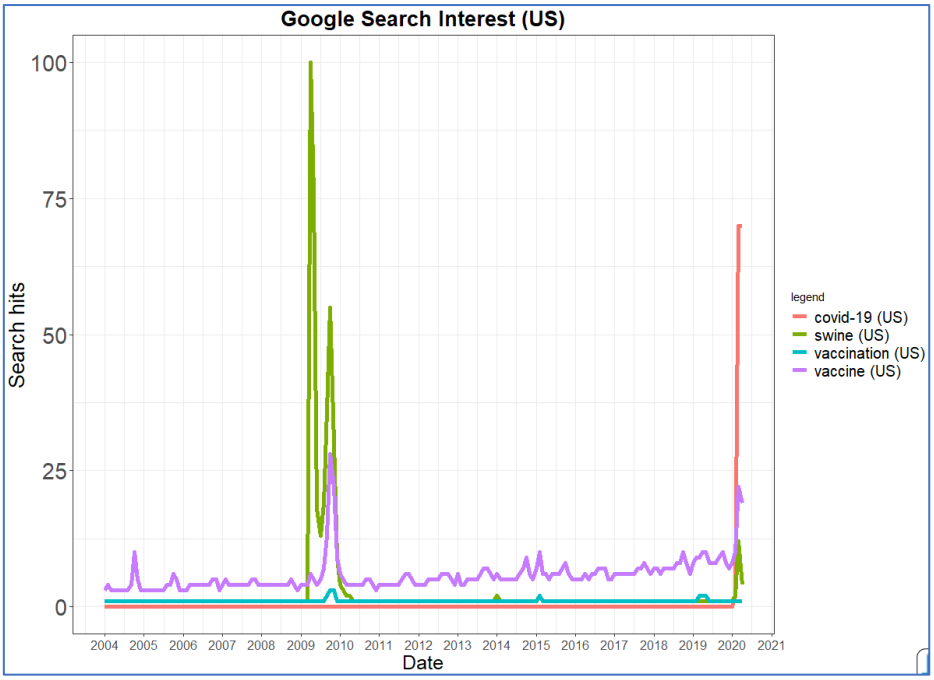

Figure 4: Google trend search volume for vaccine related keywords

4.5 Statistical Analysis and Feature Extraction of NIS Survey Data

Statistical Analysis

No null values were found in the NIS dataset. This is due to the fact that empty values were identified as "N/A" in the dataset. The rows containing "N/A" values, as well as those containing other values such as "don't know", were removed using the 'grep' function in R. Statistical tests were performed using T-Testing and Chi-Squared tests to find if there was a statistically significant effect of different variables on the outcome variable. The numerical factors (length of time child was breastfed and parent's income) were tested using T-Tests, while Chi Squared tests were used on categorical variables such as to analyse their effect on the outcome. Table 1 in the configuration manual show the results of these tests. Exploratory analysis was also performed using histograms, boxplots and the 'mice' package to test for normality and identify outliers.

The dataset was cleaned by normalizing numerical variables that were not normally distributed and or contained outliers. Plots were created using 'ggplot2' in R to find out if there was a relationship between the features from the NIS survey data and the predicted value (whether or not the child was immunized). Feature selection was also carried out using Boruta and Random Forest models to find the factors that had the most predictive power for immunization status. The underlying relationship between income and other factors such as educational attainment was also researched using the ANOVA.

Evaluation and Results:

Figure 7 shows the histograms for time spent breastfeeding and parent's income. The data was non-normal and skewed to the right. It also had outliers. The factors were normalized with the 'bbmisc' package in R. Null values were found in the length of breastfeeding column and taken out using 'na.omit' in R.

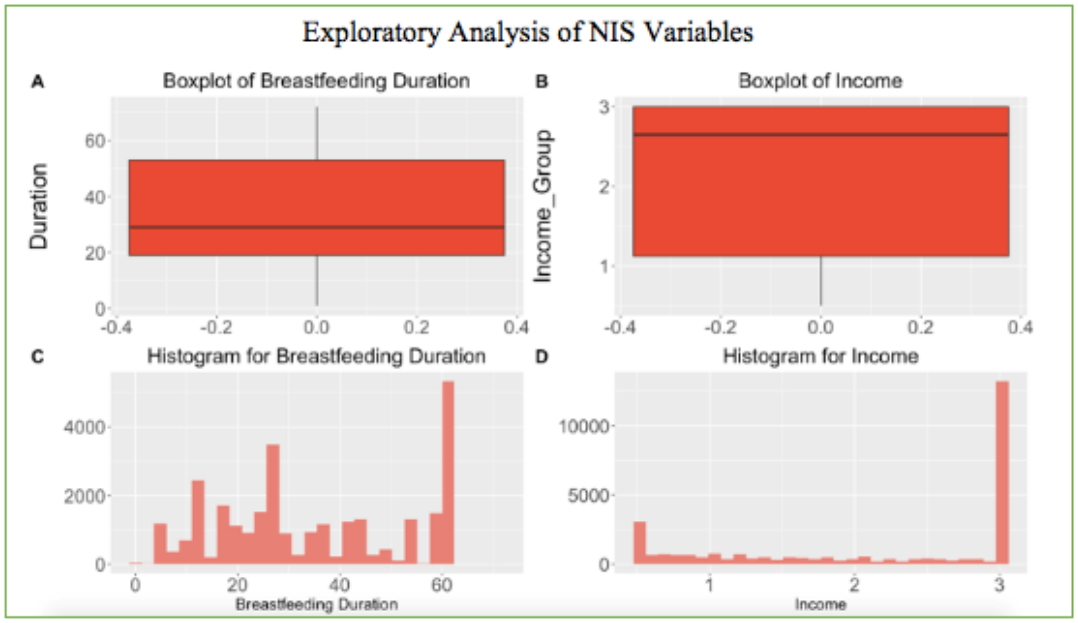

Figure 5. Boxplots and Histograms of continuous variables

Table 2 in the configuration manual shows the factors that were used in this project and if they had a statistically significant relationship with the outcome. Figure 8 shows plots that were generated using 'ggplot2' in R to visualize the relationship between some of the columns in the NIS survey with the outcome (whether or not the child was immunized).

In chart A the size of the child's family had an impact on immunization status, with larger households leading to a reduction in immunization rates, while in chart B there seems to be a slight correlation between the child number and the outcome, with the first child more likely than subsequent children to be immunized. In chart C children of parents with a higher-level education are more likely to be immunized the immunized than those with a lower level of education, with children of college graduates having the highest level of immunization. Chart D shows that children from wealthier families are more likely to be immunized than those from low income families, whilst in chart E those with private health insurance had a higher uptake of immunizations than those who use public programs and those without insurance were least likely to be immunized. In chart F racial factors have a relationship with immunization status. White children had a higher vaccine uptake, with black children having the lowest uptake.

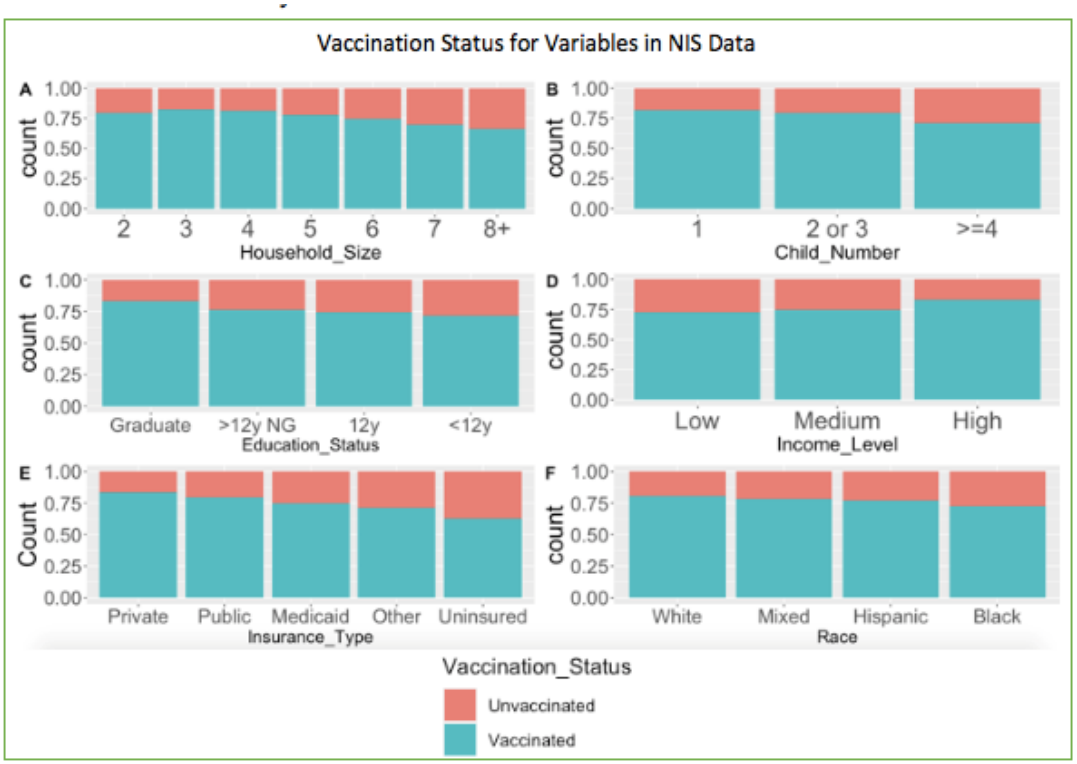

Figure 6: Vaccination Status for Variables in NIS Data

Statistical analysis was performed using Chi-Squared tests to test if there was a relationship between some of the features with income level. This was to test if the income variable may have explained some of the variation in the other variables. Family size, ethnicity, type of insurance and level of education were all found to have a statistically significant relationship with the income variable, suggesting that income may be the underlying factor explaining the effect some of the features have on immunization status.

Boruta and Random Forest models were built to generate variable importance plots to decide which features to use in the data mining models. According to Random Forest results (figure 7) the region in which the child lives has the biggest impact on immunization rates, with the length of time the child was breastfed, the child's race and the income of their parents being the next most important factors influencing immunization rates. Other factors such as whether or not the parents owned their house or if the child was in receipt of WIC (welfare) payments had a much smaller effect on immunization status. Chi-squared tests suggested that the region of the child did not affect immunization status, which may have been due to the large numbers of regions effecting the results of the chi-squared tests.

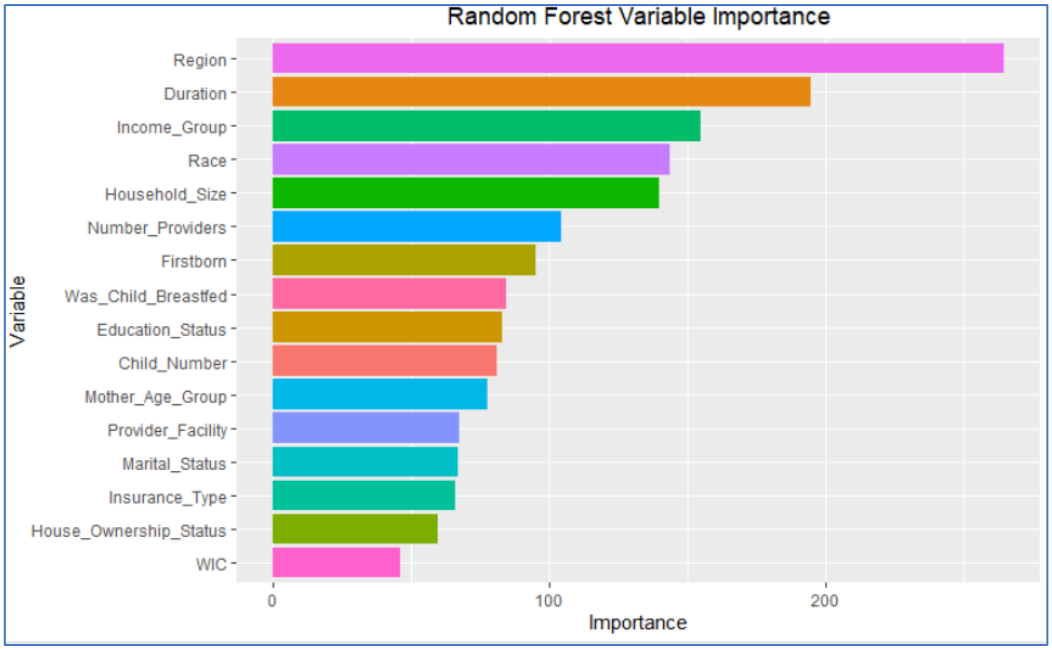

Figure 7. Variable importance as measured by Random Forest

4.6 Building Predictive Models

Ten machine learning algorithms were employed in this study to classify the immunization status of the children in the NIS survey. Each of the algorithms were first built using default parameters and then different parameters were tested to try to improve the accuracy of the models. The first stage involved creating training and testing data with an 80:20 ratio (22,959 rows in the training data and 5,740 in the test data). This chapter presents the algorithms and tuning parameters used in this project to get the best performing predictive model. Parameter tuning was performed using cross validation (CV) and grid searches (where multiple values for tuning parameters were tested).

Summary of models and tuning parameters:

Naïve Bayes: Naïve Bayes is a probabilistic that classifies data using Bayes' theorem. Parameter tuning was performed using the following values: Laplace estimator and kernel density of 0-5; using true or false values for the kernel; preprocessing the values using center, boxplot and scale techniques. The optimal values were Laplace = 1, kernel density = 1 and kernel = TRUE. The untuned model performed slightly better than the tuned model.

SVM: The SVM algorithm uses hyperplanes to classify unknown samples. There were four SVM kernels that were tested in this implementation. Kernels transform data using different methods into a suitable form for implementing SVM. The kernels implemented were 'besseldot', 'rbfdot', 'vanilla' and 'laplacedot'. The tuning parameters that were tested were: cost values of 0.0001, 00.1, 0.01, 0.1, 1 and 10; sigma = 0.001, 0.003, 0.006 and 0.009. The best lowest error came with cost = 10, each sigma value performed the same and the 'besseldot' kernel. The 'kernlab' package was used to build the models.

KNN: KNN models work by classifying unknown variables based on the values of known neighbors. KNN was tuned with K values in the range 1-20. The best performing models had

a K value of 1. The default model performed the same as the tuned model. KNN was performed using the 'class' and 'bnstruct' R modules.

Decision Trees: Decision trees are flowchart-like models that use branches to build classification models. C5.0 was used to implement these models in R. Parameter tuning was performed by testing different values for the number of trials (1, 25, 50, 75 and 100) and a Winnow status of 'True' or 'False'. The best performing models had 100 trials and the Winnow status had little impact on model performance. Tuned models performed better. Decision tress were built using the 'C50' and 'caret' modules in R.

Bagging: Bagging is a technique that combines multiple classifiers together to improve model performance and reduce overfitting. Bagging models were built with the 'ipred' package in R.

Boosting: Boosting is another ensemble learner method that uses multiple samples of the dataset to build multiple models. The model was tuned with an interaction depth ranging from 1-10 in intervals of 1, trees values of 25, 50, 75, 100 and 200 and shrinkage values of 0 or 1. The best performing model had an interaction depth of 10, 200 trees and a shrinkage value of 1.Tuned models performed best. Boosting was implemented using "caret", "xgboost" and "gbm".

Random Forest: Random Forest models generate many decision trees and combine them into an ensemble method. Tuning parameters for the random forest were: mtry value ranging from 1 to 16 in increments of 1; number of trees = 100, 250, 500 and 1000. The tuning was performed using CV and a tuning grid, and the best performaning models had mtry = 16 and 500 trees. The tuned models performed slightly better than the default model. The "randomForest" module in R was used to implement the models.

C-Forest: C-Forest algorithms are ensemble methods that combine bagging and random forest techniques. The tuning parameters for C-Forest were: mincriterion = 0.01, 0.255, 0.5, 0.745 and 0.99; maximum tree depth (MTD) = 15, 20, 25, 30, 35, 40, 45 and 50. The best performance came with a mincriterion value of 0.01 and MTD = 15. The untuned C-Forest was more accurate than the tuned C-Forest. 'caret' was used to perform C-Forest.

Artificial Neural Network: ANN models use complex neural networks to classify data. 10 CV and hyperparameter tuning were performed to find the optimal values for decay (0.1, 0.5, 0.9 and 1.3) and size (1-15). The optimal were size = 10 and decay = 0.1. untuned models performed better. Neural Networks were implemented using the "nnet" and "caret" packages.

SuperLearner: SuperLearners allow two or more data mining techniques to be combined into one ensemble learner to improve model performance. Three algorithms were combined for this study: Ranger, bagging and boosting. Other models such as ANN and SVM were attempted but not used in the final model because the run time for SuperLearners containing these algorithms was too long and wouldn't be suitable for a user-friendly platform that is intended to run in the background. Three-fold cross-validation was used to find the best performing models. The SuperLearner was run without the worst performing models (bagging), with all three models, with just the Ranger model and the final model was built with tuned parameters. The best performing model was the single SuperLearner using the Ranger algorithm. They were built using the "SuperLearner" package.

Evaluation and Results:

The performance of the machine learning algorithms are visualized using a bar plots in figure 8. The SuperLearner performed the best, with an accuracy and specificity of 76.12% and 80% and 30.01% respectively. The sensitivity for each of the models was relatively low. The SuperLearner was the best performer when all of the metrics were taken into account, which may have been due to the fact that the data contained many categorical variables, which suits tree-based models. The C50 decision tree model was the next best non-Random Forest based model in this study.

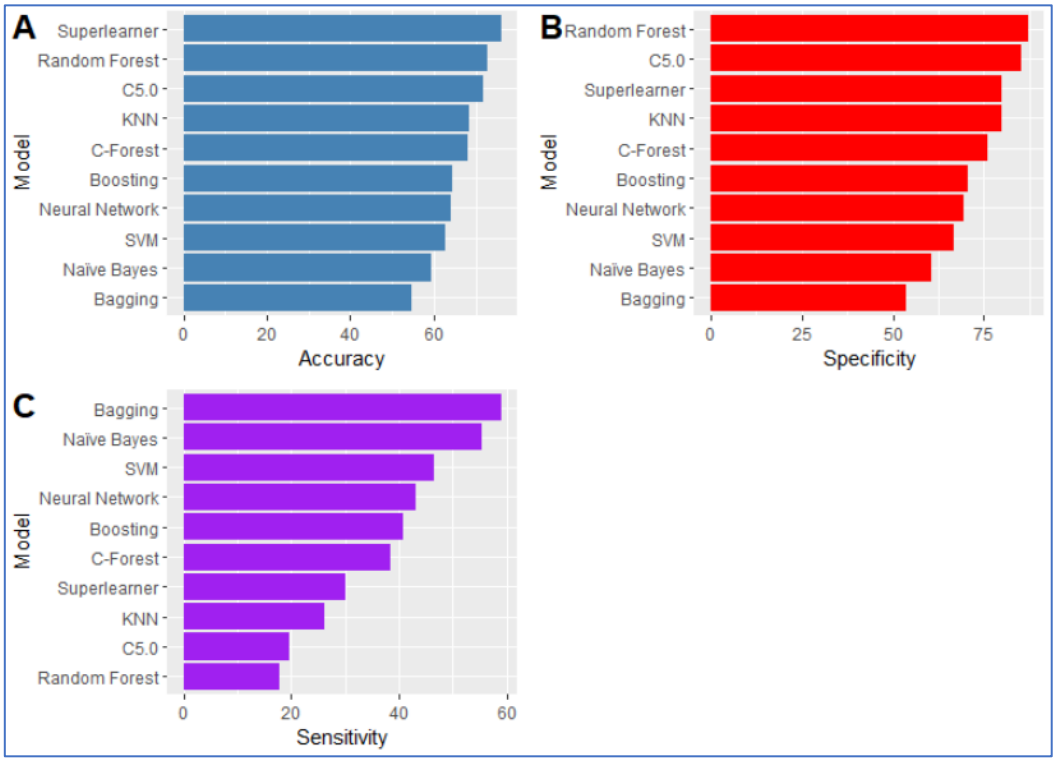

Figure 8: Comparison of Developed Models with Existing Models

The previous findings by Chandir et al. (2018) are shown in table 2. The results suggest that the study by Chandir et al. (2018) produced more accurate models. However, the platform provided by this study does have advantages over the work by Chandir et al. (2018). The data studied in this project is from multiple sources such as online content and time series data, which means that this project provides a broader understanding of attitudes towards and trends related to immunization. This study also provides a user-friendly, automated platform that doctors can use without the need for programming experience. This project provides a case study related to working with immunization data, but the study shows the advantages that programs such as R Markdown, Shiny and Flexdashboard offer to programmers in providing automated tools that can be applied to any healthcare problem.

Table 2. Comparing performance metrics with previous work

Author	Accuracy	Sensitivity	Specificity
Chandir et al.	79.1%	94.9%	54.9%
Results of this Project	76.12%	30.01%	80%

4.7 Time Series Forecasting of VRs

Implementation:

Future immunization rates were predicted using the 'ForecastML' package in R. The predicted rates for the regions of Texas, Arkansas and Oklahoma were generated based on historical values amongst children for the years 1995-2017. This data was retrieved from the CDC (CDC, 2018). The first step involved loading the data into R, then changing the format of the data to timeseries data, outliers were identified using the 'tsoutliers' package, outliers were adjusted with this package using normalization. Visualisations of the data were also generated to identify outliers. The ForecastML package was used to forecast future rates of immunization.

Results and Evaluation:

Forecasts were created for Texas and two of it's neighboring states (Arkansas and Oklahoma) (Figure 9). The results suggest that there is small increase predicted for the immunization rates in Arkansas and Oklahoma, and the values for Texas are have minor changes but the immunization rate is predicted to remain the same after the end of five years. The test for outliers showed that there were outliers in the years 2009 which were normalized using the "tsoutliers" R package.

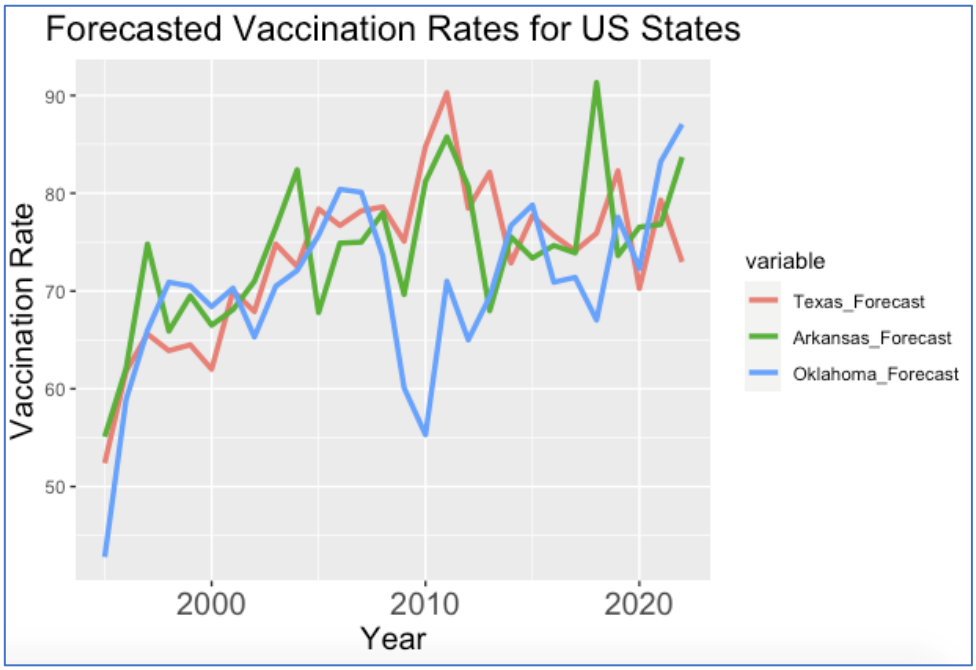

Figure 9: Timeseries forecast for vaccination rates

4.12 Persisting Data in SQL Server

Finally, the datasets from this work were stored in SSMS using dataflows in SSIS. This created a third tier in the project as a persistent layer to store to data and analyse it using SQL commands. This allows users to understand trends in the data that occur over a longer period. This is especially important with the Twitter data because TwitterR only allows users to scrape data from the previous 7 days. The implementation manual contains these SQL queries use in this study.

All implementation objectives in chapter 1, sub-section 1.2 and research question and sub-RQ have been solved (chapter 1, sub-section 1.)

5 Discussion

This project used exploratory analysis, feature selection and visuals to analyse attitudes towards immunizations, to predict immunization status in children using racial, socioeconomic and demographic factors and to predict future rates of immunizations using time series forecasting. The first part of the study analysed attitudes to immunizations using Twitter and Google Trends data. The study showed that global events such as the swine flu pandemic in 2009 and the current COVID-19 pandemic caused an increase in Google search volumes for vaccines and immunizations, and there was an interest amongst Twitter users in COVID-19 vaccines. This shows the role that online content can play in informing people about immunizations, for better or worse. Sentiment analysis showed that the two main emotions expressed by Twitter users towards vaccines were trust and fear, which highlights the polarizing nature of online discussion of the topic and the role that the medical community can play in using social media to give people information that makes them trust vaccines.

Donahue et al. (2014) found that a doctor recommending the HPV immunization increased the chance of the parent immunizing their child against HPV. Trust in immunizations may be improved by online campaigns by doctors or public healthcare workers, or by ad campaigns on search engines such as Google by groups such as the CDC or WHO.

Several factors were shown to influence the likelihood of a child being immunized such as income, educational attainment and insurance status. These findings could help in increasing immunization uptake and reducing deaths caused by infectious diseases. The relationship between insurance status and immunization status suggests that those with private health insurance are more likely to immunize their child. Lipton et al.'s (2014) research suggested that the Affordable Care Act (a policy which increased the number of people with private insurance by over 20 million people (Congressional Budget Office (2016)) improved uptake of the HPV vaccine by around 854,000 individuals. More policies that improve healthcare access, such as policies that improve access to private insurance or public healthcare, may help increase immunization uptake even further.

Educational attainment was also found to influence immunization uptake in this research, with parents that had higher levels of education more likely to immunize their children than those with lower levels of education. Previous work such as De Figueiredo et al. (2016) also found that increased levels of education amongst parents improved the chances of immunization in children. These findings suggest that policies that improve educational attainment may help to improve immunization uptake in children. This may involve increasing funding to public education, reducing the barriers to third level education by reducing interest rates on loans or increasing public funding for community colleges, or by improving standards in education using policies such as reduced classroom sizes. Income levels were also shown to influence immunization uptake, with lower income parents less likely to immunize their children than those with a higher income. The study found that there was a relationship between income level and insurance status, which suggests that lower income parents have reduced access to healthcare. Programs such as the ACA may help with this issue, as well as policies that reduce poverty rates such as increased funding for welfare programs.

Many other factors such as region of residence and household size also played a role in immunization uptake, which shows the complex nature of the problem. This study aimed to provide healthcare workers with a platform that could simplify this problem using automation.

The last stage of the project generated timeseries forecasts for future immunization coverage in three different states in the US. The findings suggest that immunization rates have increased across the three states since 1995 and the forecasts suggest that the immunization rates will continue to improve until 2022. The data showed that the rates of immunization were highly variable since 1995, and the MAE of

As well as the high variance in the data, the performance of the forecasting models could also have been negatively impacted by the small size of the timeseries dataset, which only contained 22 years of data.

6 Conclusion and Future Work

The learning outcomes for the study involved gaining knowledge in data mining techniques and statistical analysis of healthcare datasets. Experience was gained in extracting, analyzing and visualizing unstructured and structured datasets such as tweets and healthcare surveys using technologies such as R, Shiny, Flexdashboard, SSIS and SSMS. These tools were chosen because they are commonly used by data analysts and because they allow for the automation of tasks. The objectives laid out in chapter 1.3 were all achieved. A user friendly platform that could be used by professionals without coding experience was designed to generate immunization related interactive dashboards, to predict the likelihood of parents immunizing their children, and a forecasting models were implemented that could predict future rates of immunization based on historical rates.

The study also achieved the aim of creating a dashboard to visualize the data using R, Shiny and Flexdashboard, and a statistical analysis and feature selection was performed on the data to explore the underlying nature of the data. The project was automated using the R Markdown package in R, which satisfied the objective of the project to build a user friendly platform that non-programmers such as healthcare workers could use to predict trends in immunization status in children, use time series data to forecast future rates of immunization in different areas and track attitudes of online users towards immunization programs.

Future Work

There were a number of limitations of this study that may be rectified by future research. “TwitterR” was used to scrape immunization related tweets in R, but only allows the user to scrape tweets from the previous seven days. This meant that the study could only incorporate recent tweets. Future studies should be performed over many months, if not years, to allow users to model trends in attitudes towards immunizations over a longer period of time.

Clustering models were not used as part of this study due to time constraints. These models can help make inferences from data that other techniques such as classification can’t. Future research should include clustering models such as K-means clustering to analyse underlying relationships in the data that classification techniques may miss.

The CDC dataset containing immunization rates between 1995-2017 was a relatively small dataset with only 22 year of data. Other researchers should look to find larger datasets with more timepoints such as datasets with quarterly rates or datasets over longer periods of time to obtain more accurate forecasting predictions.

Future researchers could also use the platform for purposes other than analysing immunization related data. The COVID-19 pandemic has spread at different rates throughout the world and this platform could be used to map the spread of the disease using online content. Twitter posts containing keywords such as “COVID” or “Coronavirus” or Google search volumes for symptoms such as “shortness of breath” and “fever” could be used to model and visualize the areas that are likely to have infection outbreaks. Other factor such as race, income status and educational attainment could also be used as features in machine learning models to predict the likelihood of infection.

The spread of infection could also be predicted using infection rates over a period of time to forecast infection/mortality rates in the future. Machine learning models that use infection rates by area could be built to predict future rates of infections in an area and to identify areas that are most at risk of infection. This would allow public health officials to decide which locations are most likely to be hotspots of infection when deciding where to send supplies such as ventilators, drugs or immunizations. The platform would allow healthcare workers and public health officials to map the spread of the disease programmatically without the need for coding experience and would automatically generally visuals such as maps of infection rates for the analysis of the pandemic.

References

Anguera, A., et al. (2016) Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry, *Computational and Structural Biotechnology Journal* 14(-): 185-199.

Atwell, E. et al. (2013) 'Nonmedical Vaccine Exemptions and Pertussis in California, 2010' Pediatrics October 2013, 132 (4) pp.624-630;

Bhattacharya, P. et al. (2010) Automated Data Mining: An Innovative and Efficient Web-Based Approach to Maintaining Resident Case Logs. *Journal of Graduate Medical Education*. 2(4), pp. 586-570,

Blank, G. (2016) The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*. 35(6), pp.679-697.

Burton, J. (2018) 'The Most Spoken Languages In America.' World Atlas. Available at: https://www.worldatlas.com/articles/the-most-spoken-languages-in-america.html

Canavan M. et al. (2014). ‘Correlates of complete childhood vaccination in East African countries.’ *PLoS One*; 9: e95709.

CDC (2018) '1995 through 2017 Childhood Combined 5-vaccine Series Coverage Trend Report' Available at: https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/data-reports/5-series/trend/index.html

CDC (2019(a)) 'Measles Cases and Outbreaks'. Available at: https://www.cdc.gov/measles/cases-outbreaks.html

CDC (2019(b)) 'Measles Elimination.' Available at: https://www.cdc.gov/measles/elimination.html

CDC (2019(c)) 'NIS-Child Data and Documentation for 2015 to Present.' Available at: https://www.cdc.gov/vaccines/imz-managers/nis/datasets.html

Centre for Disease Control (2019(d)) National Update on Measles Cases and Outbreaks — United States, January 1–October 1, 2019.' Available at: https://www.cdc.gov/mmwr/volumes/68/wr/mm6840e2.htm

Centre for Disease Control (2020) 'Vaccines and Immunizations: Glossary.' Available at: https://www.cdc.gov/vaccines/terms/glossary.html

Chandir, S., et al. (2018) Using Predictive Analytics to Identify Children at High Risk of Defaulting From a Routine Immunization Program: Feasibility Study. *JMIR Public Health and Surveillance*. 4(3): 63.

Congressional Budget Office (2016) *Federal Subsidies for Health Insurance Coverage for People Under Age 65: 2016 to 2026.* Available at: https://www.cbo.gov/publication/51385. [Last accessed: 12 August 2020].

Crouch, E. and Dickes, L.A. (2015) A Prediction Model of Childhood Immunization Rates, *Applied Heatlh Economics and Health Policy*. 12(-), 243–251

De Figueiredo, A. (2016) Forecasted trends in vaccination coverage and correlations with socioeconomic factors: a global time-series analysis over 30 years. *The Lancet* 4(10): pp.726-35.

Donahue, K.L., et al. G.D. (2014) Acceptability of the human papillomavirus vaccine and reasons for non-vaccination among parents of adolescent sons *Vaccine*. 32(31): 3883-3885.

Foy, H. (2020) *Russia set to begin Covid-19 vaccinations within weeks*. Available at: https://www.irishtimes.com/news/world/europe/russia-set-to-begin-covid-19-vaccinations-within-weeks-1.4327421. [Last accessed 13 August 2020].

Freed, G.L., et al. (2010) Parental vaccine safety concerns in 2009. *Pediatrics.* 125 (4): 654-659.

Garrett, P.W., et al. (2010) Representations and coverage of non-English-speaking immigrants and multicultural issues in three major Australian health care publications. *Journal of the Australian Health Promotion Association*. 23(2): 84 – 85.

Hayate et al. (2016) Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction' *Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pp.76--86

Hu, Y. et al. (2014) 'Completeness and timeliness of vaccination and determinants for low and late uptake among young children in eastern China. *Human Vaccine Immunotherapy*. 10: 1408–15.

Huang, T. et al. (2015) 'Promises and challenges of big data computing in health sciences,' *Big Data Research*, 2(1), pp. 2–11, DOI: 10.1109/jbhi.2016.2636665.

Kang et al. (2017) 'Semantic network analysis of vaccine sentiment in online social media.' *Vaccine*. 35(29), pp. 3621-3638. doi: 10.1016/j.vaccine.2017.05.052

Kolodziejski, LR. (2014) Harms of Hedging in Scientific Discourse: Andrew Wakefield and the Origins of the Autism Vaccine Controversy. *Technical Communication Quarterly*. 23(3): 165-183.

Larson et al. (2016) The State of Vaccine Confidence 2016: Global Insights Through a 67-Country Survey. *EBio Medicine* –(-): 295-301.

Larson H.J. et al. (2012) Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: A systematic review of published literature, 2007–2012.' *Vaccine*. 32, pp.2150–2159.

Lipton, B.J. and Decker, S.L. (2015) ACA Provisions Associated With Increase In Percentage Of Young Adult Women Initiating And Completing The HPV Vaccine, *Health Affairs*. 43(5): 757-764.

Lu, P.J., O'Halloran, A., Williams, W.W., Lindley, M.C., Farrall, S., & Bridges, C.B. (2018) Racial and ethnic disparities in VR among adult populations in the U.S. *Vaccine.* 49(6): -.

Mavragani, A. and Ochoa, G. (2018) The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. *Big Data and Cognitive Computing.* 2(1): 2.

Mckee, C. and Bohannon, K. (2016) Exploring the Reasons Behind Parental Refusal of Vaccines, *The Journal of Pediatric Pharmacology and Therapeutics* 21(2): 104-109.

National Geographic (2016) *Herd Immunity: Strength in Numbers*. Available at: https://www.nationalgeographic.org/article/herd-immunity-strength-numbers/. [Last accessed 7 August 2020].

Palca, J. (2020) *Vaccine Candidate Delivers Protection In A Single Shot (In Monkeys).* Available At: https://www.npr.org/sections/coronavirus-live-updates/2020/07/30/897267139/vaccine-candidate-delivers-protection-in-a-single-shot-in-monkeys. [Last accessed 7 August 2020]

Pavlopoulou, I.D. (2013) Immunization coverage and predictive factors for complete and age-appropriate vaccination among preschoolers in Athens, Greece: a cross- sectional study. *BMC Public Health*. 13(-): 908.

Pourat N, et al. (2012). Role of insurance, income, and affordability in human papillomavirus vaccination. *The American Journal of Managed Care*. 18(6): 320-330.

Salathe, M. et al. (2011) Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *Plos Computational Biology*. 7(10): e1002199.

Shi H.Y. et al. (2012) Comparison of Artificial Neural Network and Logistic Regression Models for Predicting In-Hospital Mortality after Primary Liver Cancer Surgery. *PLOS ONE* 7(4): e35781.

Signorini A, et al. (2011) The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS ONE* 6(5): e19467.

Smith, P.J. et al. (2010) The association between intentional delay of vaccine administration and timely childhood VR. *Public Health Reports*, 125(4): 534-41.

Spencer A.M., et al (2014) ‘Sociodemographic factors predicting mother's cervical screening and daughter's HPV vaccination uptake,’ *Journal of Epidemiology Community Health* 68(-): 571-577.

Taksdal et al., (2013). Predictors of uptake of influenza vaccination--a survey of pregnant women in Western Australia. *Australian Family Physician*. 42(8): 582-6.

Tomeny, T.S. (2017) Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. *Social Science & Medicine*. 191(-), pp.168-175.

Tustin, J.L. (2018) Internet Exposure Associated With Canadian Parents’ Perception of Risk on Childhood Immunization: Cross-Sectional Study, *JMIR Public Health Surveillance*. 4(1) e7.

Wilson, K. and Keelan, J. (2013) Social Media and the Empowering of Opponents of Medical Technologies: The Case of Anti-Vaccinationism, *Journal of Medical Internet Research* 15(5): e103.

Xie, C., et al. (2018) User Profiling in Elderly Healthcare Services in China: Scalper Detection, *IEEE Journal of Biomedical and Health Informatics* 22(6): 1796-1806,