Project-One-Group-3

Jasmine Dogu, Brian Wimmer, Christos Chen - Group 3

01/04/2021

Contents

Research Question and Hypotheses

Goal: Through this project, our group hopes to utilize both the **Support Vector Machine** (SVM) and **Random Forest Classifier** models to predict whether a person is or is not going to get the H1N1 vaccine. These two algorithms were selected for the reasons that are provided below in the "Why Support Vector Machine and Random Forest Classifier" section.

We will first begin by explaining some background information on the H1N1 virus, the dataset, and the two machine learning algorithms. From there, we will proceed with the model creation and the parameter tuning. Lastly, we will draw conclusions from our findings and discuss future applications.

General Question: Will the SVM or the Random Forest Model Predict the Likelihood of a Person Getting the H1N1 Vaccine Better?

Null Hypothesis: The SVM Polynomial Kernel will not outperform the Random Forest Classifier Model with regards to its F1 Score **Alternative Hypothesis**: The SVM Polynomial Kernel will outperform the Random Forest Classifier Model with regards to its F1 Score

Note: The F1 score was the metric used to primarily evaluate the two models. To determine if the F1 scores are statistically significant, a **t-test** will be utilized. More information about this can be found in the "Why F1 Score" section below.

Why a Focus on F1 Score

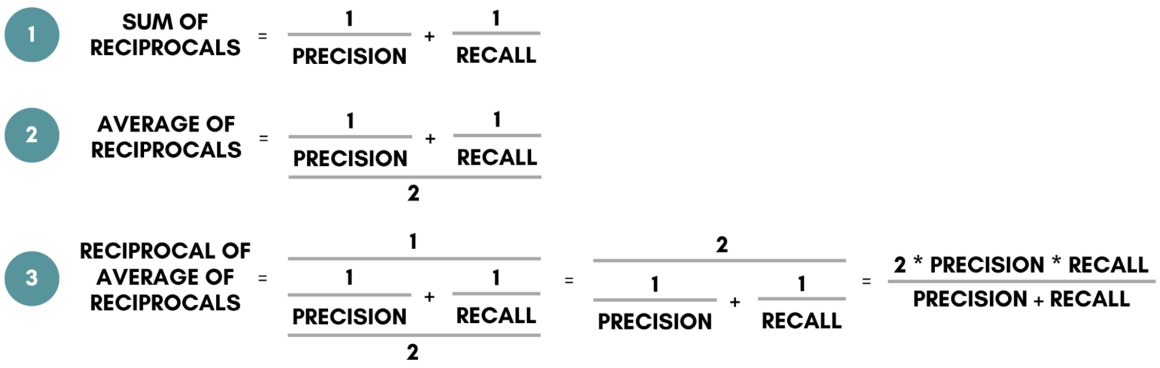


Figure 1: Reference 3

Throughout our project, we will be looking at multiple metrics to evaluate the performance of the SVM and Random Forest models. These include **accuracy, precision (specificity), and recall (sensitivity)**. While all of these metrics will play a role in how we tune the models' parameter and determine which model is more successful, to answer our question of which model outperforms the other we will be looking at the **F1 value**.

The F1 value is the harmonic mean of the sensitivity and specificity rates, and it gives a better measure of the incorrectly classified cases than the accuracy metric would. It is scaled from 0 to 1, with 1 being the best. Because there is an imbalanced class distribution in our dataset, we will be analyzing the F1 score as it is a better indicator of which model is performing better. After getting the F1 scores for both the SVM and the Random Forest models, we will be utilizing a t-test to determine if the difference between the F1 scores of the models is due to chance or is statistically significant. This will determine whether a model outperforms the other.

An explanation of why the SVM and Random Forest models were specifically selected for this project will be explained later after some data cleaning.

Background on H1N1

The flu is an illness that is caused by the influenza virus, which can lead to various symptoms that include, but are not limited to, high fevers and sore throats. The Swine Flu, in particular, is a novel influenza A (H1N1) virus that emerged in Spring 2009. It is a subtype of the Influenza A Virus and is considered an **orthomyxovirus** that contains the glycoproteins **haemagglutinin** and **neuraminidase**. The Swine Flu was initially detected in the United States, eventually spreading to the rest of the world. It is said to have "contained a blend of flu genes that hadn't been previously seen in animals or people." The Swine Flu, although very similar to the Seasonal Flu, was found to be more contagious, and less people were found to have existing resistance. (Reference 2)

About the Dataset

The selected dataset for this project is originally from the National 2009 H1N1 Flu Survey but can also be found on Driven Data.

Before building our machine learning algorithms, having a clear understanding of the data is critical. We must familiarize ourselves with the origin, size, key characteristics, behavior, and type of data. Our dataset is originally composed of **38 columns** and **26,707 observations**. Each observation accounts for one person who responded to the National 2009 H1N1 Flu Survey. There are five general groupings that each feature falls into. These data categories can be found below with a few examples of the types of variables that would belong to it:

1. Presumptions
- h1n1_concern, h1n1_knowledge
2. Behavioral: public health measure and avoidance strategies
- behavioral_face_mask, behavioral_large_gatherings, behavioral_touch_face
3. Opinion: respondent's opinion regarding the vaccine
- opinion_h1n1_risk, opinion_h1n1_sick_from_vacc
4. Demographics
- age_group, income_poverty, education, hhs_geo_region
5. Health Considerations: doctor recommendations, conditions
- doctor_recc_h1n1, chronic_med_condition, health_insurance

According to the CDC, initially there were **734,367 landline numbers** that were considered. After being narrowed down due to the unresponsiveness or the age of the respondent, a total of **105,499** respondents that were adults were found eligible. However, only **43.2%** of these individuals completed the interview. It is fair to assume that there is a **non-response bias** in play with the survey. This is important to keep in mind as it can be a source of explanation as to why the data may be imbalanced. It is very likely that individuals who were pro-vaccination were more likely to want to answer the National 2009 H1N1 Flu Survey, therefore, skewing the results. (Reference 1)

With responses from both the **adults and the children**, the CDC states that there are a total of **56,656** people who responded to the National 2009 H1N1 Flu Survey; these were composed of both landline and

cellphone interviews. In our dataset, because we only have data for **26,707** respondents, it is safe to assume that the dataset is a sample of the original National 2009 H1N1 Flu Survey. (Reference 1)

Before getting into the details of why the SVM and Random Forest models were selected, we must clean the data and get a better understanding of it.

Data Cleaning

Reading in the Dataset

The original dataset was provided in two csv files. To begin our analysis, we must first combine the two into one dataframe. This is because one csv file has the features we are interested in while the other has the labels for our observations (whether the individual received the H1N1 and Seasonal Flu Vaccine or not). We matched up each of the rows by the respondent_id, which was a column in both the features and the label csv files.

Removing the Seasonal Flu Columns

```
df <- df[, -which(names(df) %in% c("doctor_recc_seasonal","opinion_seas_vacc_effective","opinion_seas_r
head(df)
```

```
##   respondent_id h1n1_concern h1n1_knowledge behavioral_antiviral_meds
## 1             0            1              0                         0
## 2             1            3              2                         0
## 3             2            1              1                         0
## 4             3            1              1                         0
## 5             4            2              1                         0
## 6             5            3              1                         0
##   behavioral_avoidance behavioral_face_mask behavioral_wash_hands
## 1                    0                    0                     0
## 2                    1                    0                     1
## 3                    1                    0                     0
## 4                    1                    0                     1
## 5                    1                    0                     1
## 6                    1                    0                     1
##   behavioral_large_gatherings behavioral_outside_home behavioral_touch_face
## 1                           0                       1                     1
## 2                           0                       1                     1
## 3                           0                       0                     0
## 4                           1                       0                     0
## 5                           1                       0                     1
## 6                           0                       0                     1
##   doctor_recc_h1n1 chronic_med_condition child_under_6_months health_worker
## 1                0                     0                    0             0
## 2                0                     0                    0             0
## 3               NA                     1                    0             0
## 4                0                     1                    0             0
## 5                0                     0                    0             0
## 6                0                     0                    0             0
##   health_insurance opinion_h1n1_vacc_effective opinion_h1n1_risk
## 1                1                           3                1
## 2                1                           5                4
```

```
## 3               NA                           3                1
## 4               NA                           3                3
## 5               NA                           3                3
## 6               NA                           5                2
##   opinion_h1n1_sick_from_vacc     age_group        education  race    sex
## 1                           2 55 - 64 Years       < 12 Years White Female
## 2                           4 35 - 44 Years         12 Years White   Male
## 3                           1 18 - 34 Years College Graduate White   Male
## 4                           5     65+ Years         12 Years White Female
## 5                           2 45 - 54 Years     Some College White Female
## 6                           1     65+ Years         12 Years White   Male
##              income_poverty marital_status rent_or_own  employment_status
## 1             Below Poverty    Not Married         Own Not in Labor Force
## 2             Below Poverty    Not Married        Rent           Employed
## 3 <= $75,000, Above Poverty    Not Married         Own           Employed
## 4             Below Poverty    Not Married        Rent Not in Labor Force
## 5 <= $75,000, Above Poverty        Married         Own           Employed
## 6 <= $75,000, Above Poverty        Married         Own           Employed
##   hhs_geo_region               census_msa household_adults household_children
## 1       oxchjgsf                  Non-MSA                0                  0
## 2       bhuqouqj MSA, Not Principle  City                0                  0
## 3       qufhixun MSA, Not Principle  City                2                  0
## 4       lrircsnp      MSA, Principle City                0                  0
## 5       qufhixun MSA, Not Principle  City                1                  0
## 6       atmpeygn      MSA, Principle City                2                  3
##   employment_industry employment_occupation h1n1_vaccine
## 1                                                     0
## 2           pxcmvdjn              xgwztkwe            0
## 3           rucpziij              xtkaffoo            0
## 4                                                     0
## 5           wxleyezf              emcorrxb            0
## 6           saaquncn              vlluhbov            0
```

Given the recent COVID-19 pandemic, our group chose to focus only on the vaccination for H1N1, which is also known as the Swine Flu. We believe that our findings will be interesting in regards to comparing it to COVID-19. Therefore, we removed the five columns that only dealt with the Seasonal Flu.

H1N1 was still seen as a "new virus", although it was a different strain of the seasonal flu. Similarly, the COVID-19 virus and pandemic is also new and relatively unknown. We should be able to draw some similarities between skepticism and opinions regarding vaccinations between the two viruses.

Looking at the Structure of the Data: Type and NA's

```
str(df)
```

```
## 'data.frame':    26707 obs. of  33 variables:
##  $ respondent_id             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ h1n1_concern              : int  1 3 1 1 2 3 0 1 0 2 ...
##  $ h1n1_knowledge            : int  0 2 1 1 1 1 0 0 2 1 ...
##  $ behavioral_antiviral_meds : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_avoidance      : int  0 1 1 1 1 1 0 1 1 1 ...
##  $ behavioral_face_mask      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_wash_hands     : int  0 1 0 1 1 1 0 1 1 0 ...
##  $ behavioral_large_gatherings: int  0 0 0 1 1 0 0 0 1 1 ...
```

```
##  $ behavioral_outside_home    : int  1 1 0 0 0 0 0 0 1 0 ...
##  $ behavioral_touch_face      : int  1 1 0 0 1 1 0 1 1 1 ...
##  $ doctor_recc_h1n1           : int  0 0 NA 0 0 0 0 1 0 0 ...
##  $ chronic_med_condition      : int  0 0 1 1 0 0 0 1 0 1 ...
##  $ child_under_6_months       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ health_worker              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ health_insurance           : int  1 1 NA NA NA NA NA 1 NA 1 ...
##  $ opinion_h1n1_vacc_effective: int  3 5 3 3 3 5 4 5 4 4 ...
##  $ opinion_h1n1_risk          : int  1 4 1 3 3 2 1 2 1 2 ...
##  $ opinion_h1n1_sick_from_vacc: int  2 4 1 5 2 1 1 1 1 2 ...
##  $ age_group                  : chr  "55 - 64 Years" "35 - 44 Years" "18 - 34 Years" "65+ Years" ...
##  $ education                  : chr  "< 12 Years" "12 Years" "College Graduate" "12 Years" ...
##  $ race                       : chr  "White" "White" "White" "White" ...
##  $ sex                        : chr  "Female" "Male" "Male" "Female" ...
##  $ income_poverty             : chr  "Below Poverty" "Below Poverty" "<= $75,000, Above Poverty" "Be
##  $ marital_status             : chr  "Not Married" "Not Married" "Not Married" "Not Married" ...
##  $ rent_or_own                : chr  "Own" "Rent" "Own" "Rent" ...
##  $ employment_status          : chr  "Not in Labor Force" "Employed" "Employed" "Not in Labor Force"
##  $ hhs_geo_region             : chr  "oxchjgsf" "bhuqouqj" "qufhixun" "lrircsnp" ...
##  $ census_msa                 : chr  "Non-MSA" "MSA, Not Principle  City" "MSA, Not Principle  City"
##  $ household_adults           : int  0 0 2 0 1 2 0 2 1 0 ...
##  $ household_children         : int  0 0 0 0 0 3 0 0 0 0 ...
##  $ employment_industry        : chr  "" "pxcmvdjn" "rucpziij" "" ...
##  $ employment_occupation      : chr  "" "xgwztkwe" "xtkaffoo" "" ...
##  $ h1n1_vaccine               : int  0 0 0 0 0 0 0 1 0 0 ...
```

Looking at the structure of the dataset, we notice a few things. Initially, we notice that there are a few empty strings in some of the variables, for example employment_occupation and empoloyment_industry, that we want to convert to NA's. This will make it easier for us when we are analyzing the number of missing datapoints we have in order to figure out what variables we may want to consider dropping. Additionally, all of these variables must be converted to factors. This is because of the nature of the survey and types of variables we have present.

We notice that for certain variables, specifically hhs_geo_region, employment_industry, and employment_occupation, these use a classification defined by the U.S. Dept. of Health and Human Services. Because these variables are encoded for confidentiality purposes, and the encoding is not provided online, we are unable to make use these variables. Therefore, these columns will be dropped from the dataset.

Changing the Empty Strings

```
df <- replace(df, df == "", NA)
```

Here, we replaced all rows with empty strings to contain the value "NA" instead. This will come in handy as we proceed with the analysis.

Looking at Na's and Removing Specific Columns

```
colSums(is.na(df))
```

```
##              respondent_id                h1n1_concern
##                          0                          92
##             h1n1_knowledge  behavioral_antiviral_meds
##                        116                          71
```

```
##        behavioral_avoidance        behavioral_face_mask
##                         208                          19
##       behavioral_wash_hands behavioral_large_gatherings
##                          42                          87
##     behavioral_outside_home       behavioral_touch_face
##                          82                         128
##           doctor_recc_h1n1       chronic_med_condition
##                        2160                         971
##        child_under_6_months              health_worker
##                         820                         804
##           health_insurance opinion_h1n1_vacc_effective
##                       12274                         391
##          opinion_h1n1_risk opinion_h1n1_sick_from_vacc
##                         388                         395
##                   age_group                   education
##                           0                        1407
##                        race                         sex
##                           0                           0
##             income_poverty              marital_status
##                        4423                        1408
##                rent_or_own           employment_status
##                        2042                        1463
##             hhs_geo_region                  census_msa
##                           0                           0
##           household_adults          household_children
##                         249                         249
##        employment_industry       employment_occupation
##                       13330                       13470
##               h1n1_vaccine
##                           0
```

Looking at the number of NA's we have present in each column, we can see that the health_insurance, employment_industry, and employment_occupation make up the top 3 columns with the highest amount of NA's.

We decided to remove the health_insurance column for a few reasons. It was the main source of NA's within the dataset (12,274), and deleting rows associated with this column would have effectively eliminated nearly half of the dataset. In addition, we did some further research into the financial obligations regarding the H1N1 vaccine. According to the Centers for Disease Control (CDC), the government wanted to avoid any economic obstacles for everyday Americans when it came to obtaining a vaccine. Vaccination providers, such as clinics or drugstores, were not allowed to charge volunteers for the vaccine, as the supplies had already been purchased by the US Government. This allowed us to conclude that health insurance, therefore, would not be as necessary or as big of an obstacle for obtaining the H1N1 vaccine, in comparison to others. Further information can be found here.

We also chose to remove the employment_industry and employment_occupation variables. Similarly to health_insurance, removing observations with "NA" for these two variables would have resulted in nearly half of our dataset being eliminated. These variables include "codes" that are correlated to US Census data regarding industry and occupation types. Converting these to factor variables and then taking the average or median to engineer data for the NA's would be inaccurate. We would be left with very high numbers of one particular industry and occupation, which would not help the model in determining if people would get an H1N1 vaccine or not.

Lastly, we removed the respondent_id at this step. This variable was important when we were matching the features to the labels. However, for the machine learning algorithm, it is not needed.

```r
df <- df[, -which(names(df) %in% c("employment_occupation","health_insurance","employment_industry", "hh
df<- na.omit(df)
#write.csv(df,"df2.csv", row.names = TRUE) #saving to a new csv file
df2<- read.csv("df2.csv")
nrow(df)
```

```
## [1] 19656
```

Here, along with removing the top 3 columns with the highest number of NA's, we also removed the hhs_geo_region variable for reasons mentioned above. We are left with 19,656 observations. Because this is still a significant portion of the original number of observations the dataset had (approximately 73.6%), we proceed with our analysis.

In this step of our project, we decided to save the current dataframe to a new csv file. This is because in the next step we will be coercing all of the features to factors; therefore, it is a good idea to keep a copy of the clean data that contains all of the features' data types as they originally were.

Coercing the Variables to Factors

```r
df[,] <- lapply(df[,], factor)  ## as.factor() could also be used
str(df)
```

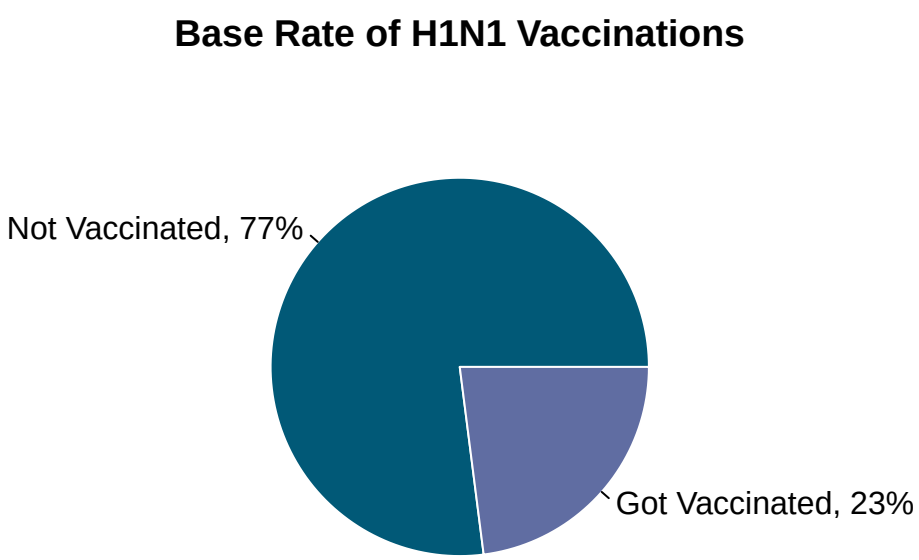
```
## 'data.frame':    19656 obs. of  28 variables:
##  $ h1n1_concern               : Factor w/ 4 levels "0","1","2","3": 2 4 2 3 4 1 2 1 3 3 ...
##  $ h1n1_knowledge             : Factor w/ 3 levels "0","1","2": 1 3 2 2 2 1 1 3 2 2 ...
##  $ behavioral_antiviral_meds  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_avoidance       : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 2 2 ...
##  $ behavioral_face_mask       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_wash_hands      : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 1 2 ...
##  $ behavioral_large_gatherings: Factor w/ 2 levels "0","1": 1 1 2 2 1 1 1 2 2 2 ...
##  $ behavioral_outside_home    : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 2 1 1 ...
##  $ behavioral_touch_face      : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 2 2 1 ...
##  $ doctor_recc_h1n1           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ chronic_med_condition      : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 2 2 ...
##  $ child_under_6_months       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ health_worker              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ opinion_h1n1_vacc_effective: Factor w/ 5 levels "1","2","3","4",..: 3 5 3 3 5 4 5 4 4 4 ...
##  $ opinion_h1n1_risk          : Factor w/ 5 levels "1","2","3","4",..: 1 4 3 3 2 1 2 1 2 1 ...
##  $ opinion_h1n1_sick_from_vacc: Factor w/ 5 levels "1","2","3","4",..: 2 4 5 2 1 1 1 1 2 2 ...
##  $ age_group                  : Factor w/ 5 levels "18 - 34 Years",..: 4 2 5 3 5 4 3 3 4 3 ...
##  $ education                  : Factor w/ 4 levels "< 12 Years","12 Years",..: 1 2 2 4 2 1 4 3 2 2 .
##  $ race                       : Factor w/ 4 levels "Black","Hispanic",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ sex                        : Factor w/ 2 levels "Female","Male": 1 2 1 1 2 2 1 2 2 2 ...
##  $ income_poverty             : Factor w/ 3 levels "<= $75,000, Above Poverty",..: 3 3 3 1 1 1 1 2 1
##  $ marital_status             : Factor w/ 2 levels "Married","Not Married": 2 2 2 1 1 2 1 1 2 1 ...
##  $ rent_or_own                : Factor w/ 2 levels "Own","Rent": 1 2 2 1 1 1 1 1 1 2 ...
##  $ employment_status          : Factor w/ 3 levels "Employed","Not in Labor Force",..: 2 1 2 1 1 1 1
##  $ census_msa                 : Factor w/ 3 levels "MSA, Not Principle  City",..: 3 1 2 1 2 1 3 1 1 1
##  $ household_adults           : Factor w/ 4 levels "0","1","2","3": 1 1 1 2 3 1 3 2 1 3 ...
##  $ household_children         : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 4 1 1 1 1 1 ...
##  $ h1n1_vaccine               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:7051] 3 17 25 27 30 32 39 40 43 45 ...
##   ..- attr(*, "names")= chr [1:7051] "3" "17" "25" "27" ...
```

```
#write.csv(df,"cleaned_df.csv", row.names = TRUE) #saving to a new csv file
```

After this step, we can verify that all of the variables are now factors by using the str() function. Because they are all factors, we can carry on with our analysis.

Why Support Vector Machine and Random Forest Classifier

```
library(dplyr)
library(tidyverse)
pie( c( nrow(df %>% filter(h1n1_vaccine==0)), nrow(df %>% filter(h1n1_vaccine==1))), labels = c("Not Va
```



After a preliminary analysis of the data topology, our group decided to compare the SVM and Random Forest models. We were curious to how these two models would perform against one another due to the nature of the algorithms. Both are **supervised machine learning algorithms** which can be used for **classification** or **regression** analysis. For the purposes of our analysis, we will be using these algorithms to classify which group a person, given various features, is likely to belong to (vaccine/no vaccine). We selected these algorithms because they are both great for imbalanced datasets, which as we can see above, our dataset has a base rate of 23% for the group who received the vaccine; because our dataset is imbalanced, it is critical to utilize algorithms that can deal with it fairly well. Similarly, both of the algorithms reduce the risk of over-fitting in their own manner. This is also extremely important for our model as overfitting to the dataset can falsely skew the accuracy of the model.

In theory, the Random Forest model would have a few benefits over the SVM model. The Random Forest is non-parametric, so outliers would not be an issue for the algorithm. Similarly, they are extremely easy to build (compared to SVMs) and are fast/scalable. Unlike with SVMs, Random Forests do not require a lot of parameter tuning. These are all characteristics of the algorithm that make it extremely favorable for this research study.

However, the SVM model also has a few general benefits over the Random Forest model. Usually, the SVM model is known to be highly efficient and accurate. Similarly, the SVM model scales well to high dimensional data, which is not significantly relevant for our dataset as we have more observations than we do features. However, the main reason why we predict that the SVM model will outperform the Random Forest model is because of the kernels that the SVM model uses. With an appropriate kernel and degree of the kernel, we can expect the algorithm to work extremely well even if the data is not linearly separable in the base feature space, which in this instance applies to our dataset.

In order to begin building our model, we now proceed to the EDA, followed by the feature selection. These are two sections that will help us gain a better understanding of the dataset and re-adjust our hypotheses, if needed.

EDA

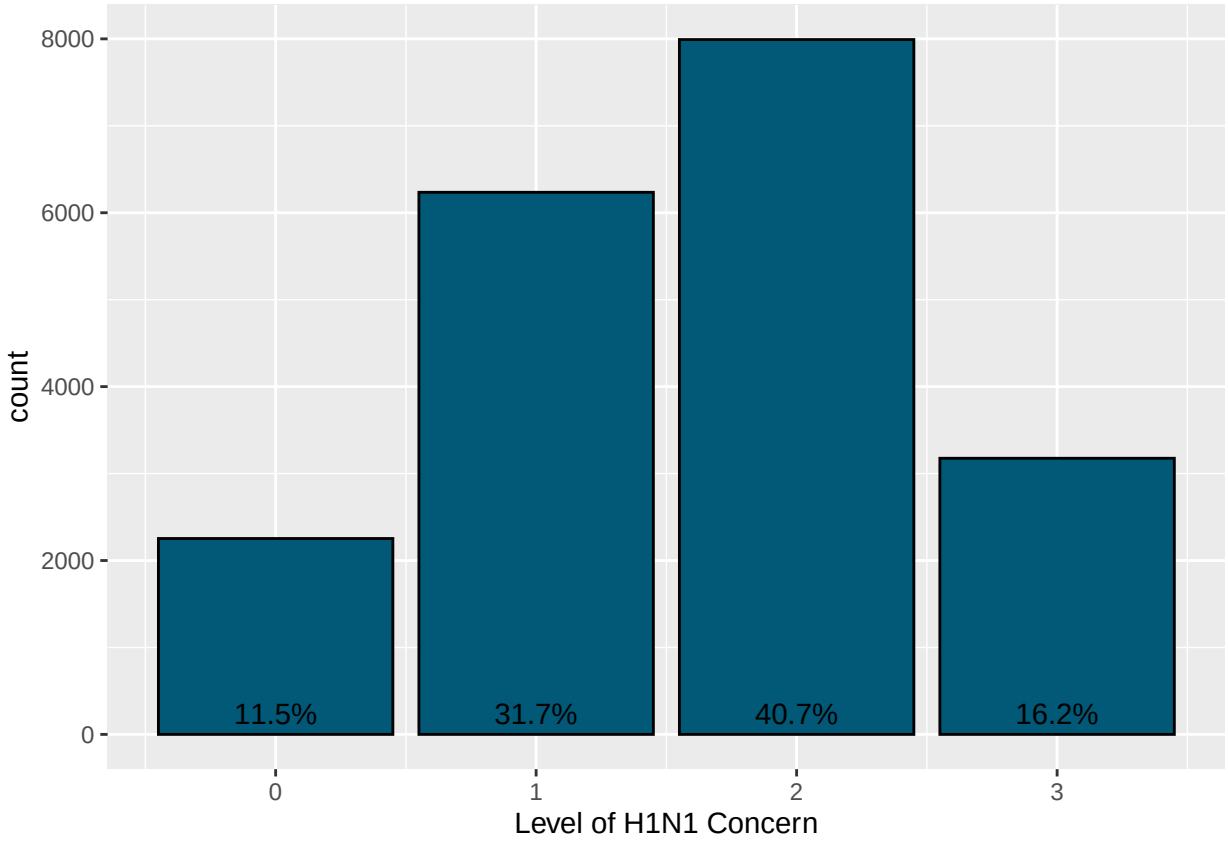
In order to continue with the project, we first wanted to explore the data and to try to get an idea of the types of people within the dataset that may choose to get the vaccine. For this, we decided to create histograms depicting different features from each of the categories we had: presumptions, behavioral, opinion, demographic, and health considerations.

Histogram for "Presumptions" Variables

```
pre_1 <- ggplot(df2, aes(x=h1n1_concern)) +
  geom_histogram(stat="count",bins=4,fill="#015977",color="black")+xlab("Level of H1N1 Concern") + geom_
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
pre_1
```



Over 70 percent of respondents presented a moderate level of concern regarding H1N1. 11.5% and 16.2% indicated "not at all concerned" and "very concerned", respectively. We would expect those that indicated "very concerned" to be the most likely to get the H1N1 vaccine.

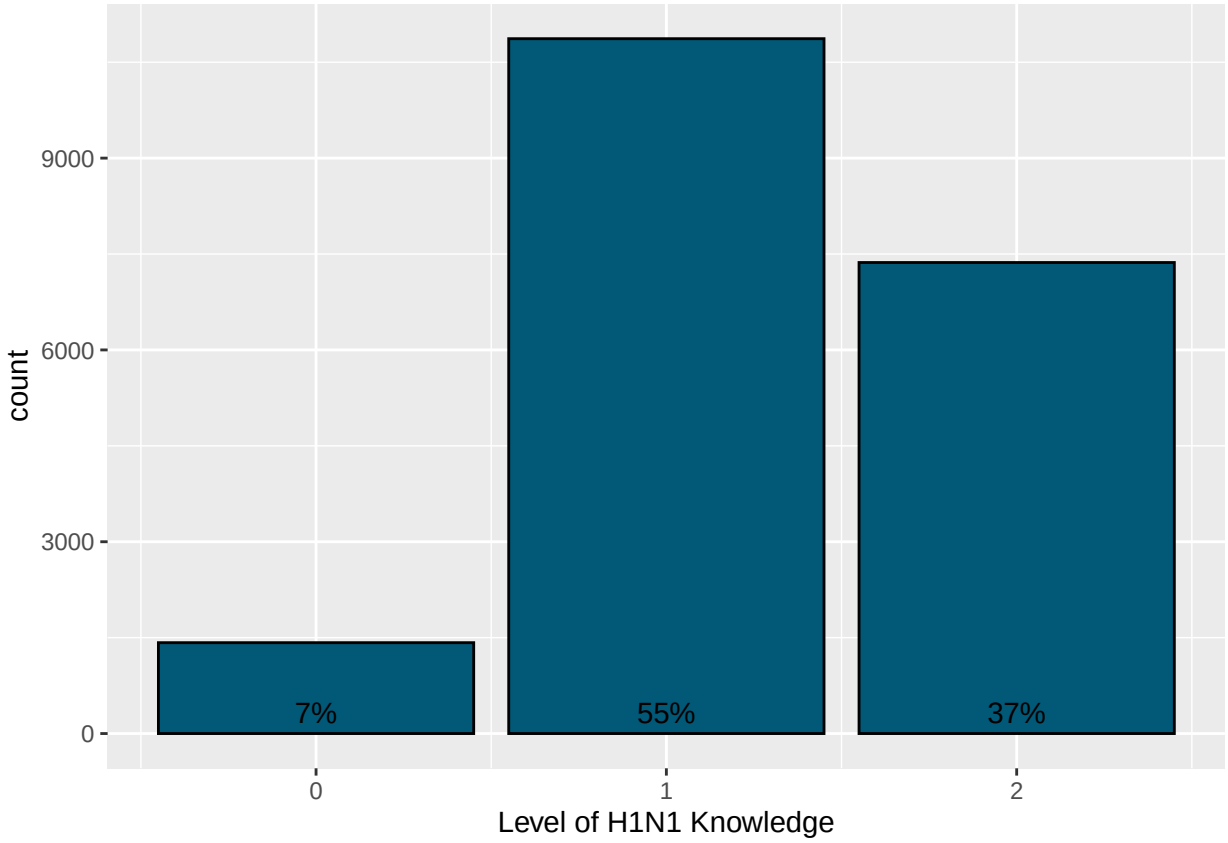
0-Not at all concerned
1-Not very concerned

2-Somewhat concerned
3-Very concerned

```
pre_2 <- ggplot(df2, aes(x=h1n1_knowledge)) +
  geom_histogram(stat="count",bins=3,fill="#015977",color="black")+xlab("Level of H1N1 Knowledge") + ge
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
pre_2
```



The histogram shows there to be a suprisingly large proportion of respondents that indicated "A lot of knowledge" of the H1N1 Virus, with 37%. Only 7% indicated "No knowledge". We would expect the majority of those that reported having "A lot of knowledge" to want the vaccine, when compared to those reporting "No knowledge".

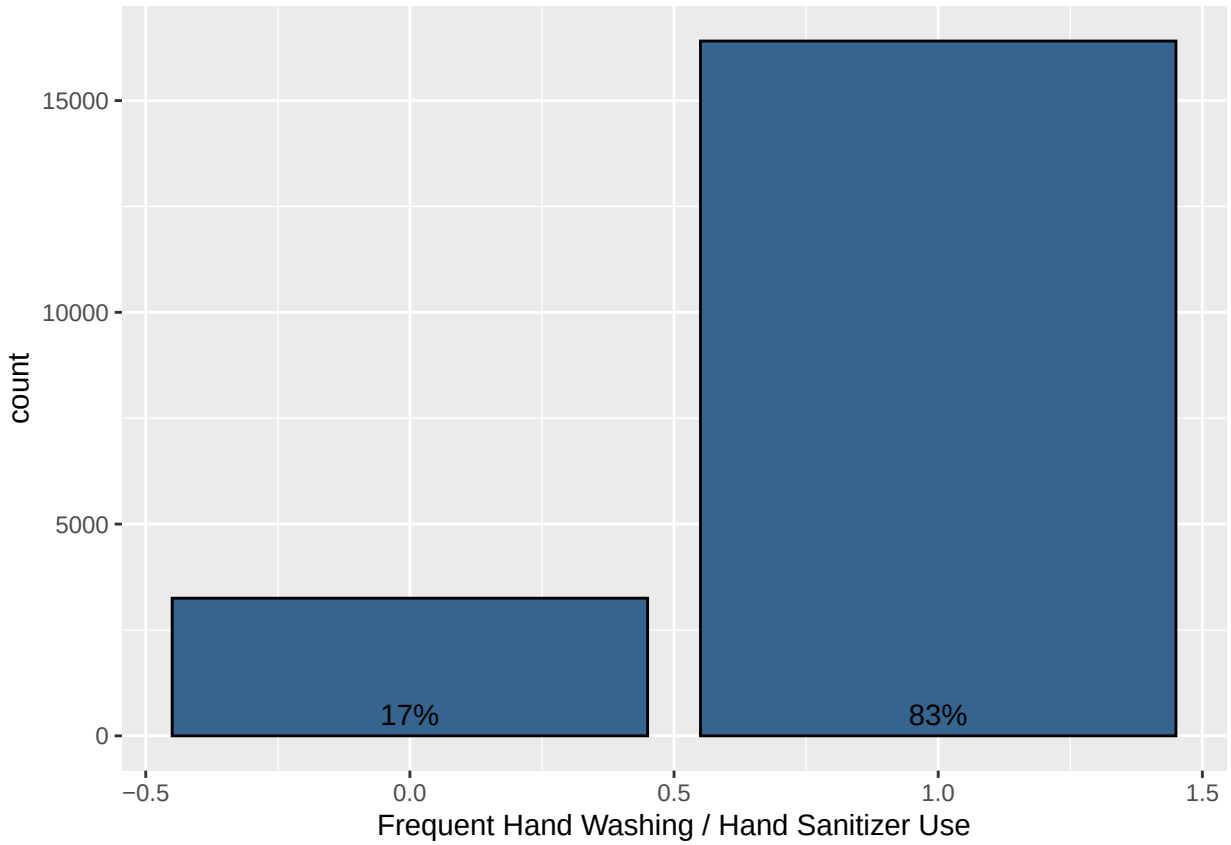
0-No knowledge
1-A little knowledge
2-A lot of knowledge

Histogram for "Behavioral" Variables

```
beh1 <- ggplot(df2, aes(x=behavioral_wash_hands)) +
  geom_histogram(stat="count",bins=2,fill="#35648F",color="black")+xlab("Frequent Hand Washing / Hand Sa
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
beh1
```



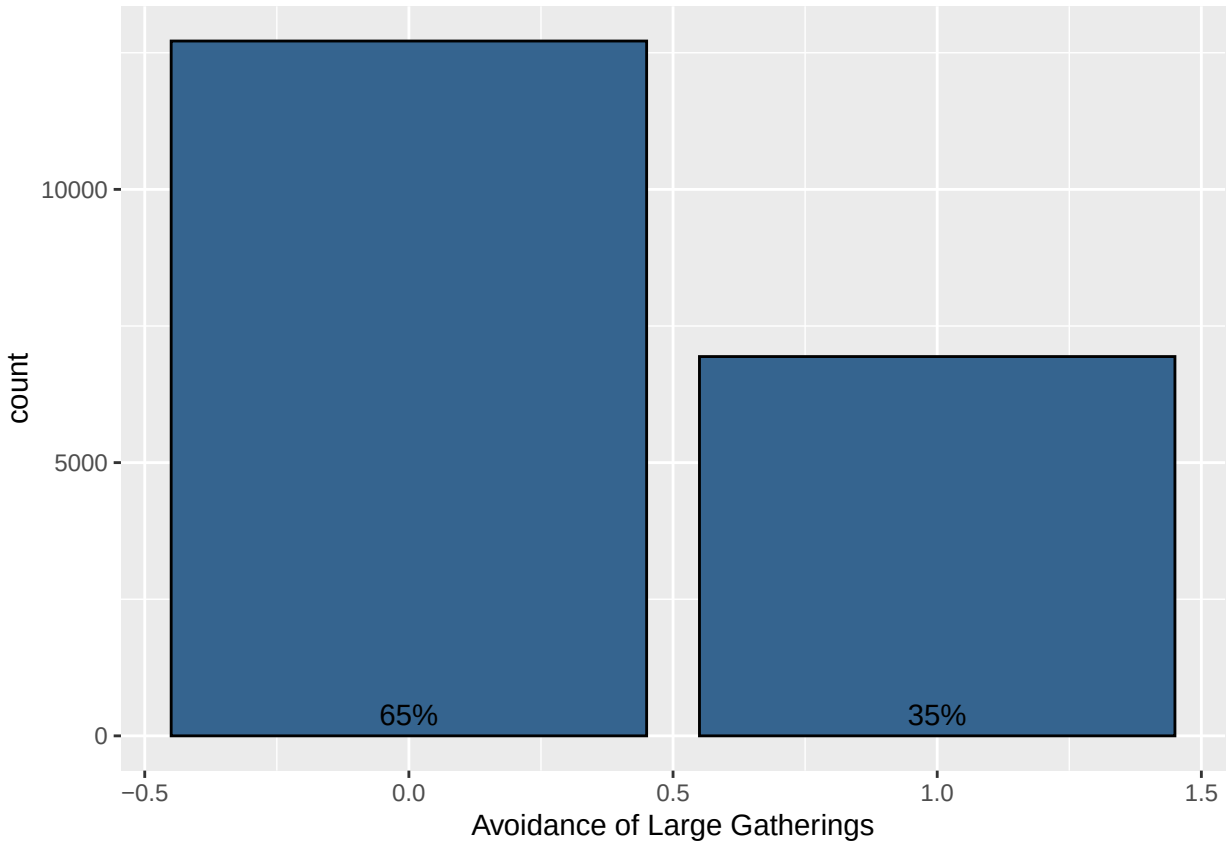
17% of respondents indicated infrequent hand washing and lack of sanitizer use, while 83% indicated the opposite. We would expect those that reported frequent usage to be more likely to get the H1N1 vaccine, as this indicates more self-awareness of healthy habits and the presence of germs.

0-No
1-Yes

```
beh2 <- ggplot(df2, aes(x=behavioral_large_gatherings)) +
  geom_histogram(stat="count",bins=2,fill="#35648F",color="black")+xlab("Avoidance of Large Gatherings")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
beh2
```



65% of respondents indicated NOT avoiding large gatherings during the H1N1 crisis, while 35% indicated the opposite. We would expect those that indicated that they practiced avoidance measures to be more likely to get the H1N1 vaccine, as they would seem to be more worried of the risks of the virus.

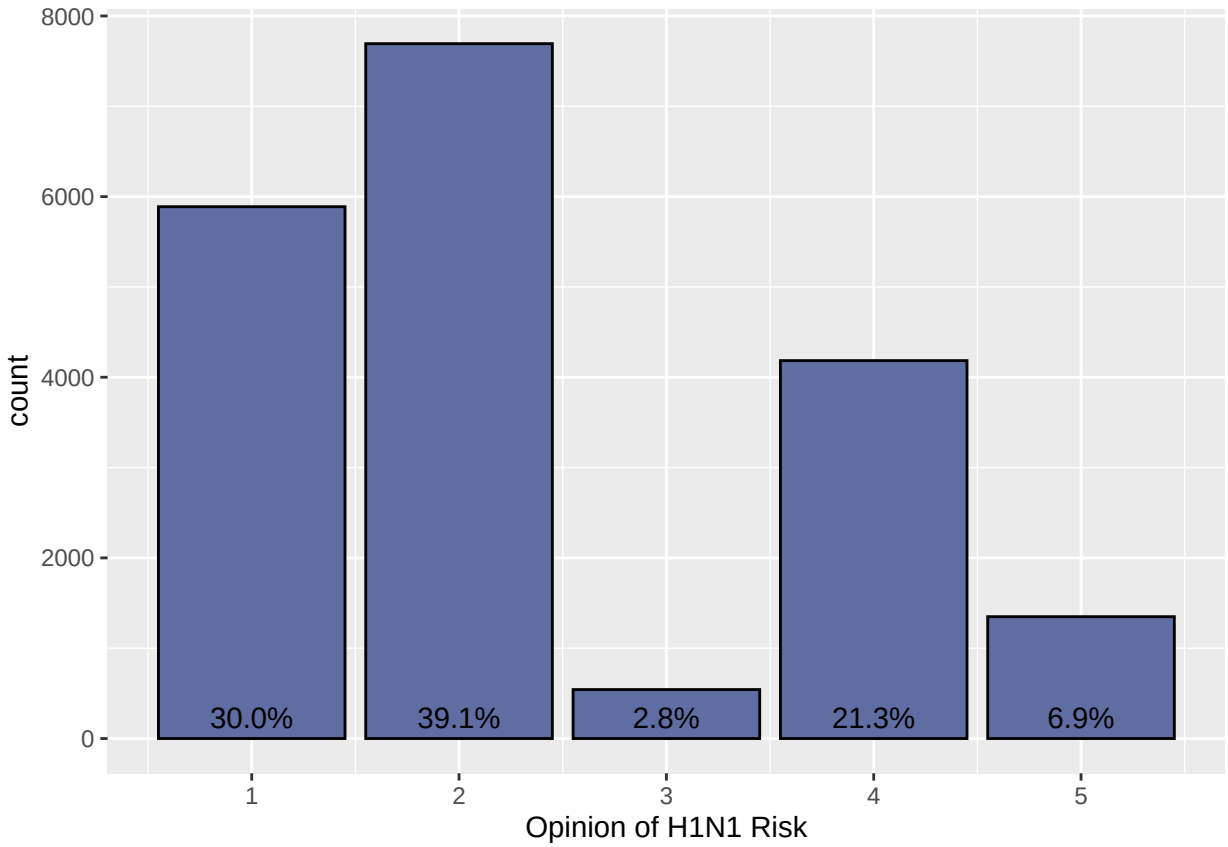
0-No
1-Yes

Histogram for "Opinion" Variables

```
opi1 <- ggplot(df2, aes(x=opinion_h1n1_risk)) +
  geom_histogram(stat="count",bins=5,fill="#606DA2",color="black")+xlab("Opinion of H1N1 Risk") + geom_
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
opi1
```



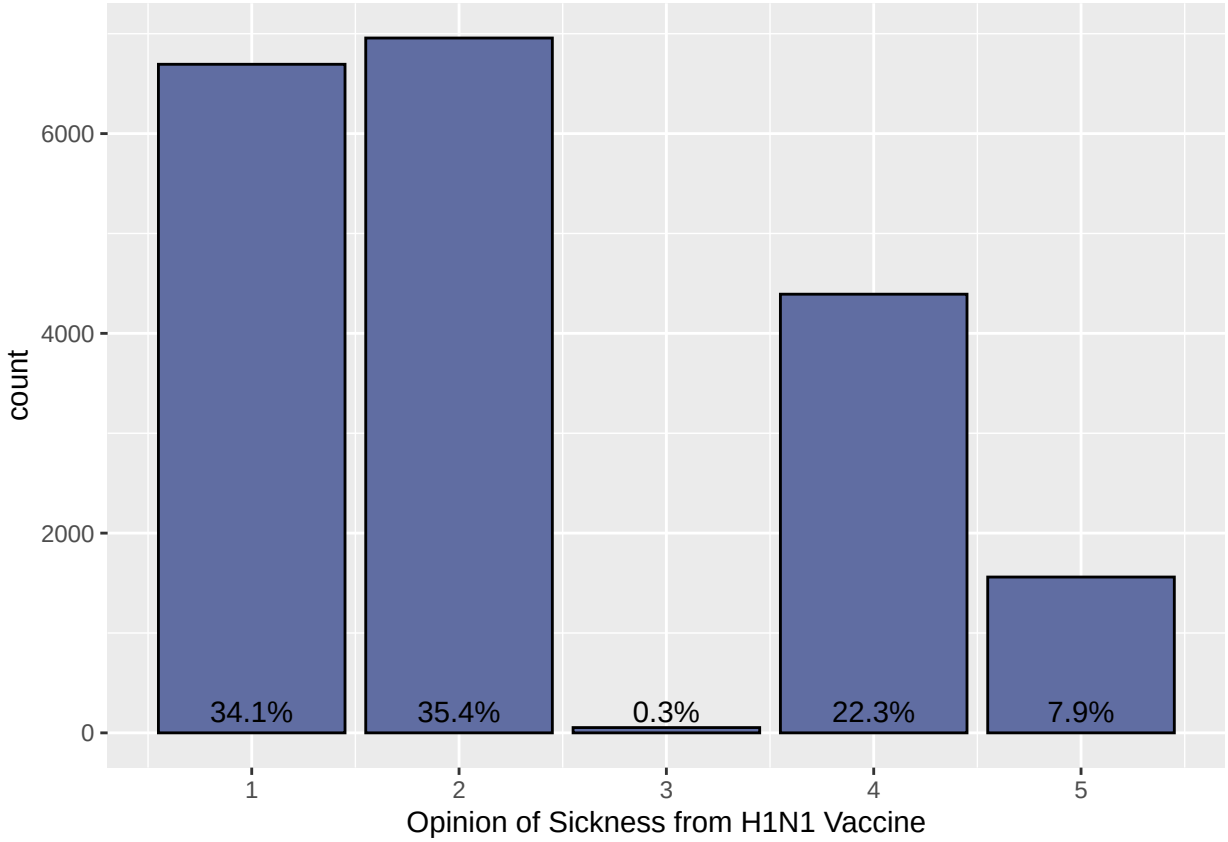
There is a relatively large disparity between opinions regarding H1N1 risk. Nearly 70% of respondents indicated the risk to be "Very Low" or "Somewhat Low", while only 28.2% indicated the risk posed by H1N1 to be "Somewhat High" or "Very High".

1-Very Low
2-Somewhat Low
3-Don't Know
4-Somewhat High
5-Very High

```
opi2 <- ggplot(df2, aes(x=opinion_h1n1_sick_from_vacc)) +
  geom_histogram(stat="count",bins=5,fill="#606DA2",color="black")+xlab("Opinion of Sickness from H1N1 V
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
opi2
```



There is a relatively large disparity between opinions regarding getting sick from the H1N1 vaccince. 69.5% of respondents indicated the risk of sickness to be "Not at all worried" or "Somewhat Worried", while only 30.2% indicated the risk of sickness to be "Somewhat Worried" or "Very Worried".

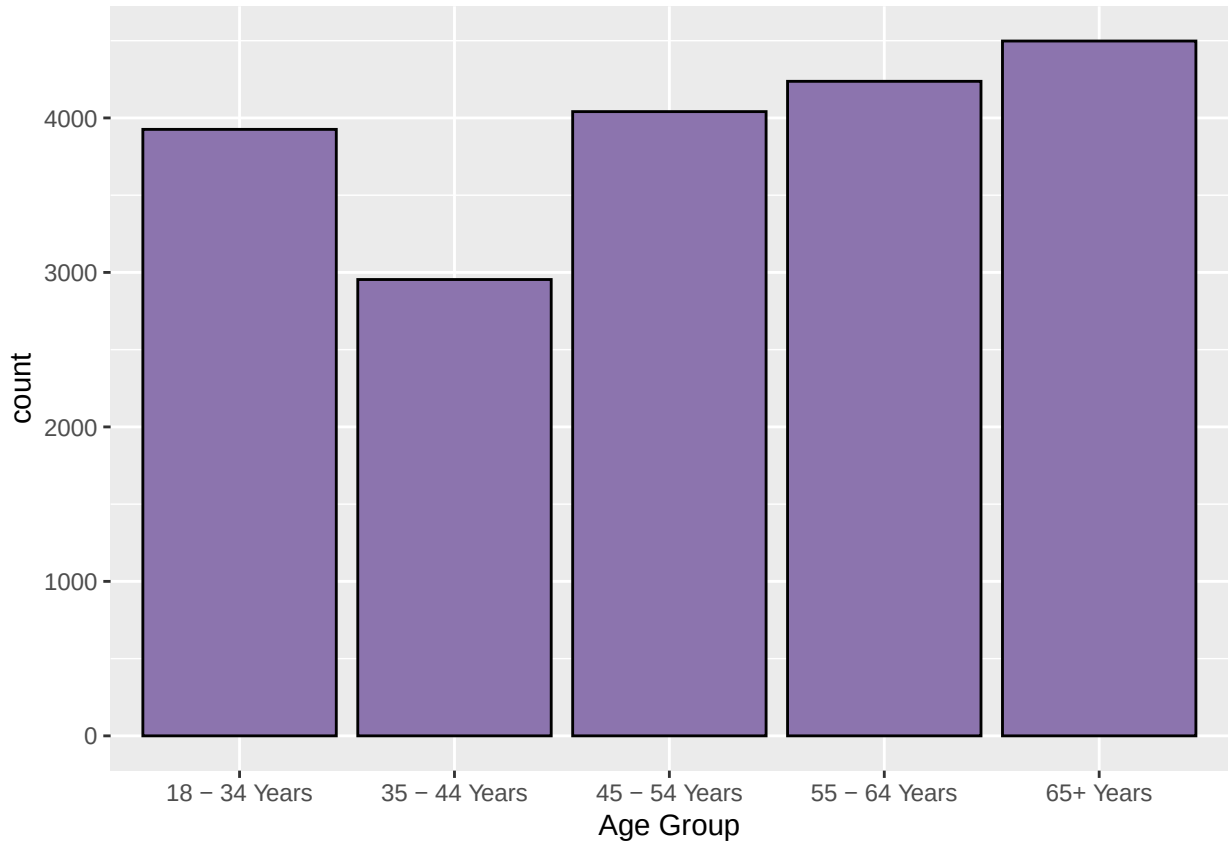
1-Not at all worried
2-Not very worried
3-Don't know
4-Somewhat worried
5-Very worried

Histogram for "Demographics" Variables

```
dem1 <- ggplot(df2, aes(x=age_group)) +
  geom_histogram(stat="count",fill="#8C74AE",color="black")+xlab("Age Group")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
dem1
```

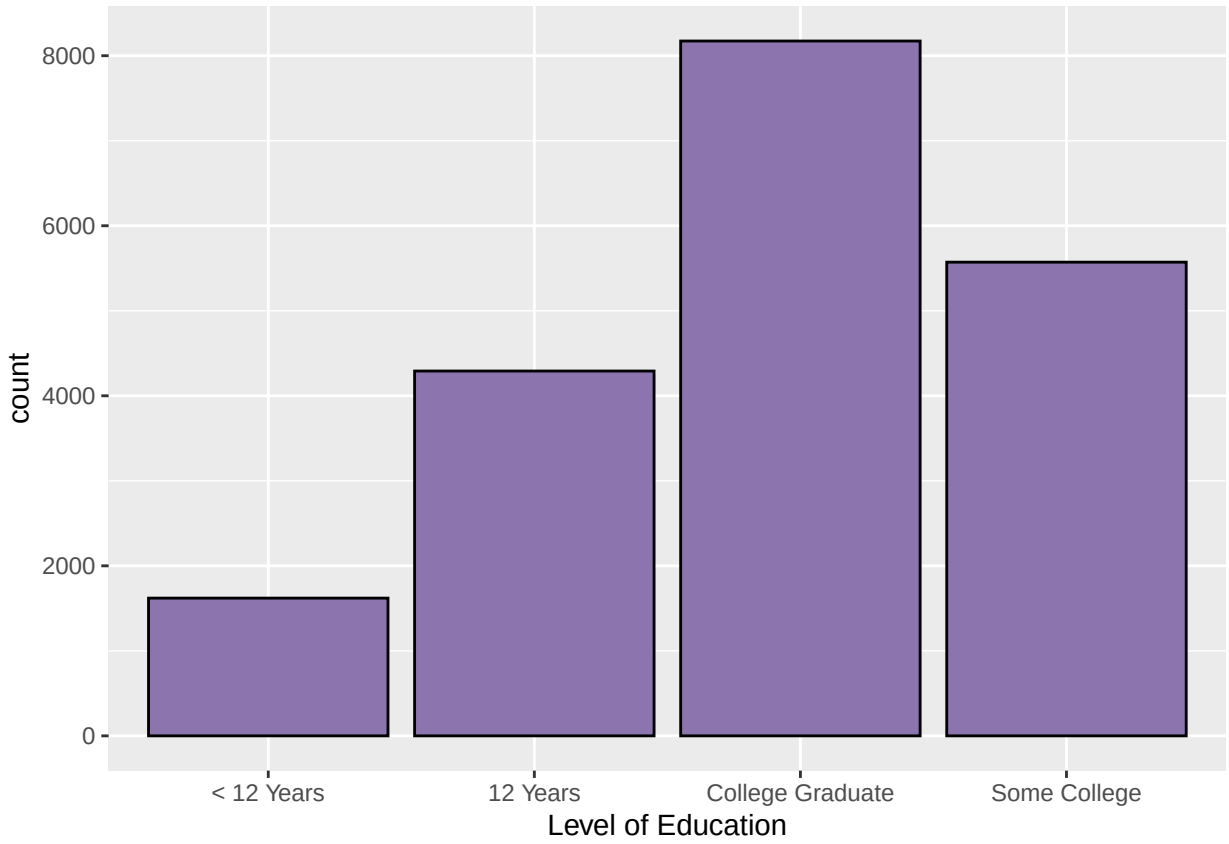


The 65+ age group was the most represented among respondents, while the 35-44 was the least represented.

```
dem2 <- ggplot(df2, aes(x=education)) +
  geom_histogram(stat="count",fill="#8C74AE",color="black")+xlab("Level of Education")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
dem2
```

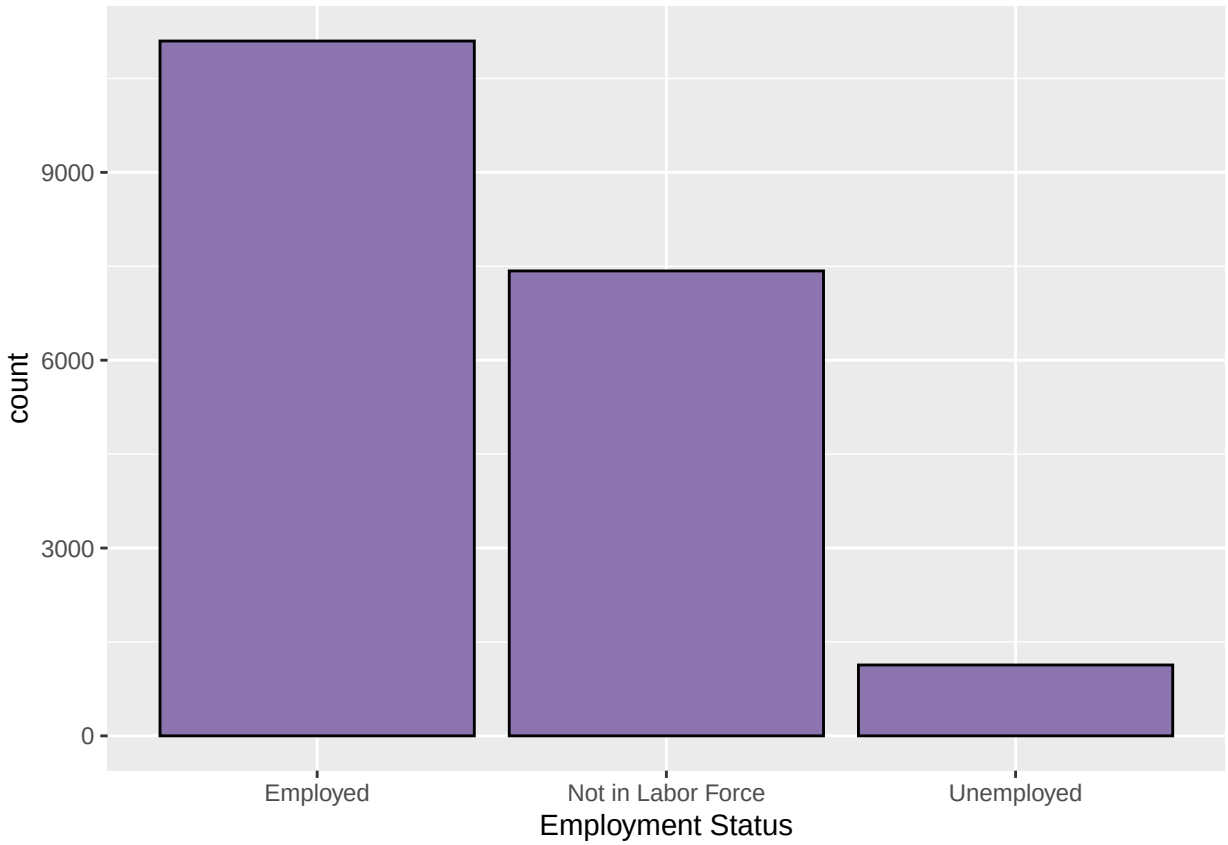


The majority of respondents indicated having at least some college. The most represented group were College Graduates.

```
dem3 <- ggplot(df2, aes(x=employment_status)) +
  geom_histogram(stat="count",fill="#8C74AE",color="black")+xlab("Employment Status")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
dem3
```



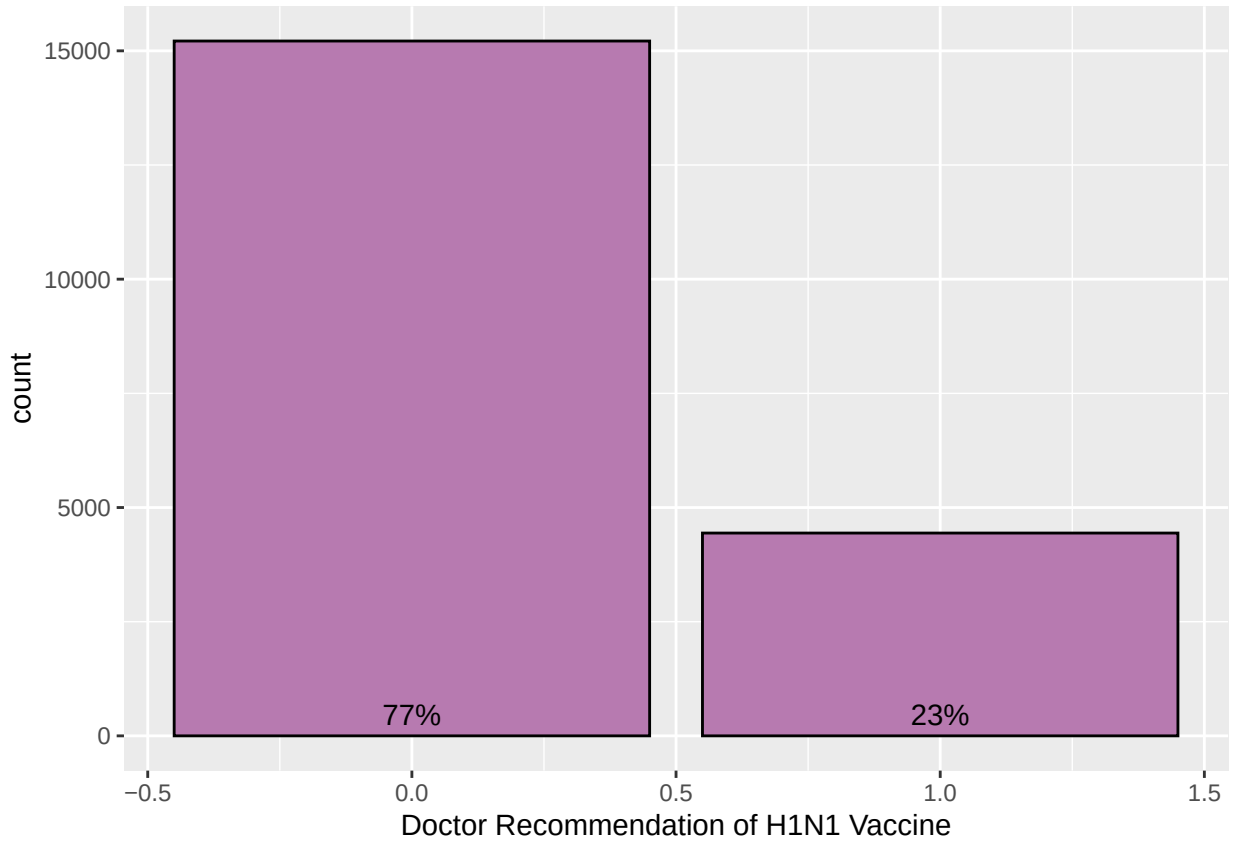
Over 50% of respondents indicated that they were employed.

Histogram for "Health Considerations" Variables

```
hc1 <- ggplot(df2, aes(x=doctor_recc_h1n1)) +
  geom_histogram(stat="count",bins=2,fill="#B77AB0",color="black")+xlab("Doctor Recommendation of H1N1 V
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
hc1
```



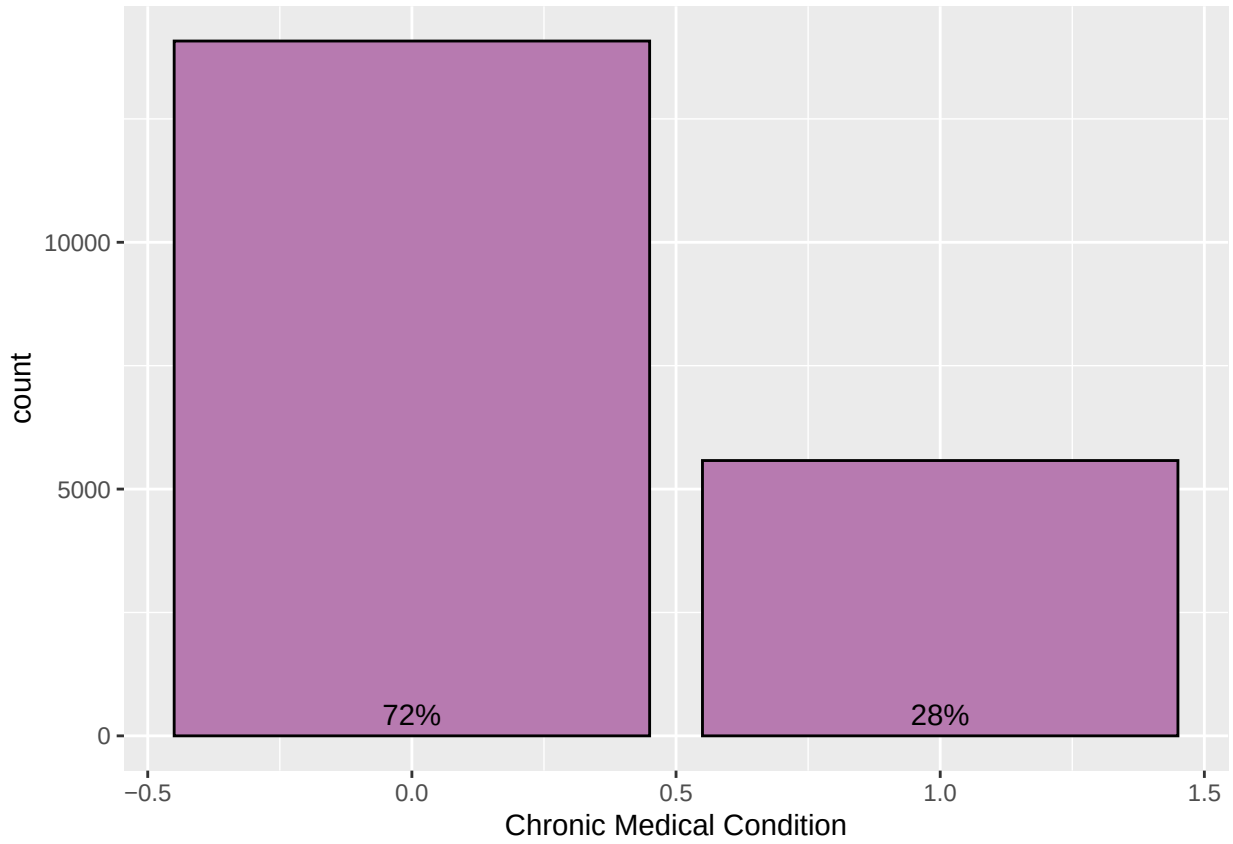
77% of respondents indicated that their doctor did NOT recommend getting the H1N1 Vaccine, while only 23% indicated the presence of a recommendation.

0-No
1-Yes

```
hc2 <- ggplot(df2, aes(x=chronic_med_condition)) +
  geom_histogram(stat="count",bins=2,fill="#B77AB0",color="black")+xlab("Chronic Medical Condition") + g
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
hc2
```



72% or respondents indicated NOT having a chronic medical condition, while 28% indicated the presence of a chronic medical condition. 0-No
1-Yes

Feature Selection

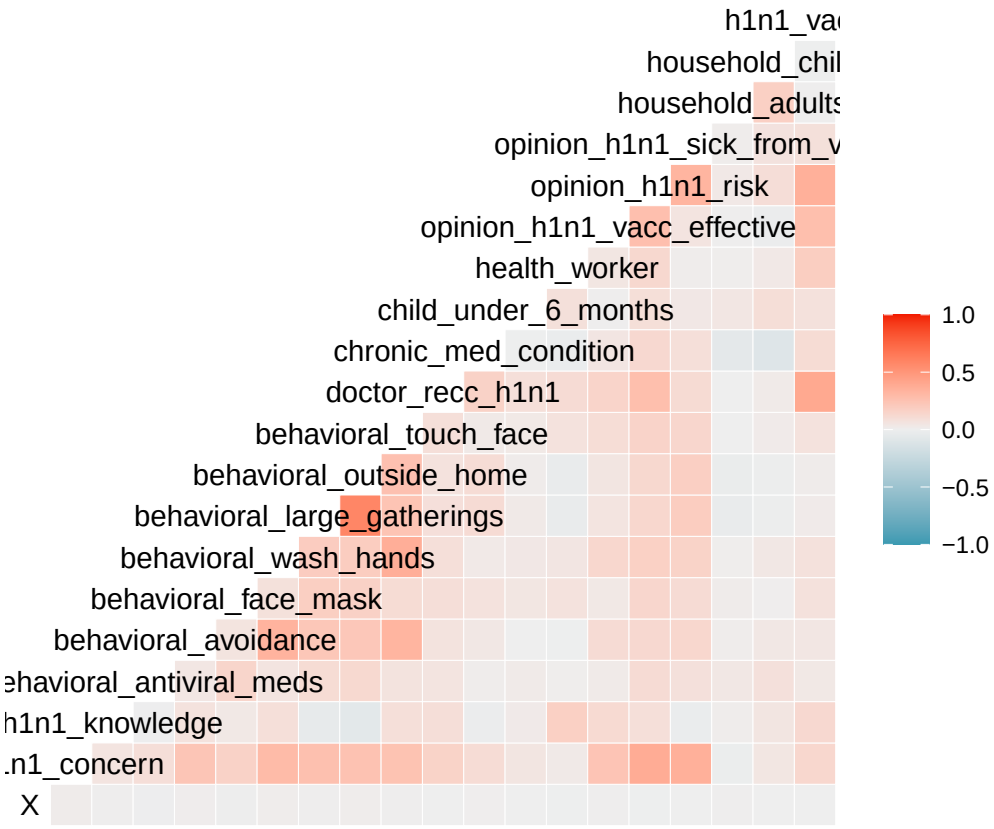
Corellogram

```
h1n1_correlation = ggcorr(df2, method = c("everything", "pearson"), title = "H1N1 Factors Correlogram")
```

```
## Warning in ggcorr(df2, method = c("everything", "pearson"), title = "H1N1
## Factors Correlogram"): data in column(s) 'age_group', 'education', 'race',
## 'sex', 'income_poverty', 'marital_status', 'rent_or_own', 'employment_status',
## 'census_msa' are not numeric and were ignored
```

```
## Warning: Ignoring unknown parameters: title
```

```
h1n1_correlation
```



A correlogram was utilized to inform and guide future efforts for feature selection. We aimed to identify variables weakly and strongly correlated to the target variable and then test model performance with & without those variables. As demonstrated above, none of the variables are particularly strongly correlated with whether someone relieved the h1n1 vaccination. However, the most correlated are **doctor recommendation**, **opinion about the risk of getting sick with the flu vaccine**, and their **opinion of whether or not the h1n1 vaccine was effective** with correlations of **0.13**, **0.11**, and **0.09** respectively.

Because we determined that no variables were highly correlated, we will not have to remove any of the features and can move on to the LASSO Regression.

LASSO Regression Model

Why LASSO Regression

Initially, we want to utilize lasso regression to reduce the feature space. LASSO regression is an analysis method that utilizes both variable selection and regularization to further enhance the predictive accuracy and interpretability. It imposes a constraint on the model parameters that causes the regression coefficients for some of the variables to shrink towards zero.

Preparing the Data

```
# Split the data into training and test set
set.seed(03092000)

training.samples <- df2$h1n1_vaccine %>%
  createDataPartition(p = 0.8, list = FALSE)
```

```
train.data  <- df2[training.samples, ]
test.data <- df2[-training.samples, ]
# Dummy code categorical predictor variables
x <- model.matrix(h1n1_vaccine~., train.data)[,-c(1,29)]
# Convert the outcome (class) to a numerical variable
y <- ifelse(train.data$h1n1_vaccine == "pos", 1, 0)
```

Computing Penalized Logistic Regression

```
#library(glmnet)
#glmnet(x, y, family = "binomial", alpha = 1, lambda = NULL)
```

Future Direction for Project

Utilizing both the polynomial kernel and the radial kernel, we will explore tuning the hyperparameters and comparing the performance of each as they both possessed similar RMSE values. We initially ran multiple kernels and selected the Polynomial Kernel due to its low RMSE of 0.401, with the Radial Kernel following close behind at an RSME of 0.417. However, because these kernel's performed very similarly with regard to our loss function, we would still like to further assess both kernels.

Following the execution of the SVM, we would like to proceed with the Random Forest and aim to tune the hyperparameters for optimal performance. Having tuned both of the models, the statistical significance of the resulting F1 Scores of each algorithm will be assessed for statistical significance with a t-test.

SVM Model

Splitting Data into Training and Testing Set

```
# Splitting the dataset into the Training set and Test set
#install.packages('caTools')
set.seed(03092000)
library(caTools)

# Creating a 70/30 split
splitting_data <- sample(1:nrow(df),
                round(0.7 * nrow(df), 0),
                replace = FALSE)

splitting_data_test <- sample(1:nrow(df),
                round(0.15 * nrow(df), 0),
                replace = FALSE)

#Creating the train and test data
svm_train <- df[splitting_data, ] #Should contain 70% of data points
svm_test <- df[splitting_data_test, ] #Should contain 15% of data points

#Checking to ensure steps above were done correctly
size_of_training <- nrow(svm_train)
```

```
size_of_total <- nrow(df)
size_of_test <- nrow(svm_test)

#Verification
paste("The Training Set contains", toString(round(size_of_training/size_of_total,2)*100), "% of the tota
```

```
## [1] "The Training Set contains 70 % of the total data"
```

```
paste("The Testing Set contains", toString(round(size_of_test/size_of_total,2)*100), "% of the total da
```

```
## [1] "The Testing Set contains 15 % of the total data"
```

Creating the Model

```
#install.packages('e1071')
library(e1071)

classifier <- svm(formula = h1n1_vaccine ~ .,
                  data = svm_train,
                  type = 'C-classification',  #Default
                  #can change degree
                  kernel = 'polynomial')  #The kernel used in training and predicting
```

Predicting the Test Set Results with Model

```
# Predicting the test set results
head(df)
```

```
##   h1n1_concern h1n1_knowledge behavioral_antiviral_meds behavioral_avoidance
## 1            1              0                         0                    0
## 2            3              2                         0                    1
## 4            1              1                         0                    1
## 5            2              1                         0                    1
## 6            3              1                         0                    1
## 7            0              0                         0                    0
##   behavioral_face_mask behavioral_wash_hands behavioral_large_gatherings
## 1                    0                     0                           0
## 2                    0                     1                           0
## 4                    0                     1                           1
## 5                    0                     1                           1
## 6                    0                     1                           0
## 7                    0                     0                           0
##   behavioral_outside_home behavioral_touch_face doctor_recc_h1n1
## 1                       1                     1               0
## 2                       1                     1               0
## 4                       0                     0               0
## 5                       0                     1               0
## 6                       0                     1               0
## 7                       0                     0               0
##   chronic_med_condition child_under_6_months health_worker
## 1                     0                    0             0
## 2                     0                    0             0
```

```
## 4                     1                   0            0
## 5                     0                   0            0
## 6                     0                   0            0
## 7                     0                   0            0
##   opinion_h1n1_vacc_effective opinion_h1n1_risk opinion_h1n1_sick_from_vacc
## 1                           3                 1                           2
## 2                           5                 4                           4
## 4                           3                 3                           5
## 5                           3                 3                           2
## 6                           5                 2                           1
## 7                           4                 1                           1
##       age_group    education  race    sex            income_poverty
## 1 55 - 64 Years   < 12 Years White Female             Below Poverty
## 2 35 - 44 Years     12 Years White   Male             Below Poverty
## 4     65+ Years     12 Years White Female             Below Poverty
## 5 45 - 54 Years Some College White Female <= $75,000, Above Poverty
## 6     65+ Years     12 Years White   Male <= $75,000, Above Poverty
## 7 55 - 64 Years   < 12 Years White   Male <= $75,000, Above Poverty
##   marital_status rent_or_own  employment_status               census_msa
## 1    Not Married         Own Not in Labor Force                  Non-MSA
## 2    Not Married        Rent           Employed MSA, Not Principle  City
## 4    Not Married        Rent Not in Labor Force      MSA, Principle City
## 5        Married         Own           Employed MSA, Not Principle  City
## 6        Married         Own           Employed      MSA, Principle City
## 7    Not Married         Own           Employed MSA, Not Principle  City
##   household_adults household_children h1n1_vaccine
## 1                0                  0            0
## 2                0                  0            0
## 4                0                  0            0
## 5                1                  0            0
## 6                2                  3            0
## 7                0                  0            0
```

```
y_pred <- predict(classifier, newdata = svm_test[-28])
```

Creating Confusion Matrix

```
# Making a Confusion Matrix
#install.packages("caret")
library(caret)
cm <- confusionMatrix(svm_test$h1n1_vaccine,y_pred, positive = "1")
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2261    6
##          1  620   61
##
##                Accuracy : 0.7877
##                  95% CI : (0.7724, 0.8023)
##     No Information Rate : 0.9773
```

```
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.127
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.91045
##             Specificity : 0.78480
##          Pos Pred Value : 0.08957
##          Neg Pred Value : 0.99735
##              Prevalence : 0.02273
##          Detection Rate : 0.02069
##    Detection Prevalence : 0.23100
##       Balanced Accuracy : 0.84762
##
##        'Positive' Class : 1
##
```

References

1. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NIS/nhfs/nhfspuf_DUG.PDF
2. https://www.cdc.gov/h1n1flu/surveillanceqa.htm
3. https://towardsdatascience.com/simplifying-precision-recall-and-other-evaluation-metrics-d066b527c6bb