12-2021

Factors Influencing Intent to Take a COVID-19 Test in the United States

Sheila Rutto
The University of Texas Rio Grande Valley

FACTORS INFLUENCING INTENT TO TAKE A COVID-19 TEST IN THE UNITED STATES.

A Thesis

by

SHEILA RUTTO

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Applied Statistics and Data Science

The University of Texas Rio Grande Valley
December 2021

FACTORS INFLUENCING INTENT TO TAKE A COVID-19 TEST IN THE UNITED STATES.

A Thesis
by
SHEILA RUTTO

COMMITTEE MEMBERS

Dr. Hansapani Rodrigo
Chair of Committee

Dr.George Yanev
Committee Member

Dr. Tamer Oraby
Committee Member

Dr. Zhijun Qiao
Committee Member

December 2021

ABSTRACT

Rutto, Sheila, Factors Influencing Intent to Take a COVID-19 Test in the United States. . Master of Science (MS), December, 2021, 60 pp., 9 tables, 17 figures, references, 36 titles.

In 2020, COVID-19 became the first pandemic in the world's history that brought the entire world to an abrupt and unexpected halt. Since the first reported case of the disease to date, the novel coronavirus has been able to wreak havoc in literary every corner of the globe and left an ever-growing number of unprecedented fatalities. The normal way of life has been disrupted, and the level of uncertainty about the end of this pandemic continues to manifest to many. Due to the urgency to bring this pandemic under control, medical officers have been able to recommend actions that people need to undertake voluntarily to assist in slowing down the spread of the disease. This study has a particular focus on COVID-19 testing as an essential measure being used in monitoring and controlling the spread of the virus. The study investigates some of the essential factors that can predict whether a person has higher odds of taking a coronavirus test or not. As it is evident, the fight against the spread of coronavirus is a collective responsibility that requires socially responsible behavior from people. This study used a portion of the data collected by the Understanding American Study (UAS) in their national longitudinal survey of the attitudes and behaviors around COVID-19 in the USA. The participants of this survey were randomly sampled using the Address-Based Sampling (ABS) from postal records drawn from across the country. The targeted sample size was 8,900 participants, but only 6,067 of them could complete the questionnaire, and this study utilized 440 of the completed responses that had no missing values. Both descriptive and inferential statistics were computed. For the descriptive analysis, frequencies were obtained as the majority of the variables were categorical. The bivariate analysis was performed using Chi-square

and Wilcoxon Sign tests. Further analysis was performed using machine learning models including classification and regression decision trees, gradient boosting(CART), random forest, and artificial neural networks, followed by logistic regression models. The findings showed significant higher odds ratios for, persons with black ethnicity (OR: 3.26, 95% CI:1.36,7.86), and obtaining news about COVID-19 from physicians (OR:4.17, 95% CI: 2.13, 8.30) towards taking a coronavirus test. Every one unit increase in the Age has shown 3% lower odds (OR:0.97, 95% CI: 0.95, 0.99) while obtaining coronavirus news from social media (OR:0.41, 95% CI:0.18, 0.88), never feeling unable to control life during the pandemic (OR:0.28, 95%CI:0.12, 0.63) and sometimes feeling unable to control life during the pandemic (OR:0.30, 95%CI:0.10, 0.87) also had shown lower odds of an individual taking part in COVID-19 testing. Random Forest (RF) model had yielded the optimal average area under the receiver operating characteristic curve value (AUC) of 78.15(SD:1.05) with an accuracy of 76.34%(SD:2.13) followed by the (CART) Decision Tree Model with an average AUC of 60.91%(SD:4.51) and an accuracy of 72.39(SD:2.89). The SHAP analysis based on the optimal RF model reveals that use of social media to obtain coronavirus information, feeling of things not going your way during pandemic, constant worrying, age and feeling unable to control life situation during the pandemic were found to be the most influencing factors. The study recommends that the health care authorities consider these factors when conducting their awareness programs on the importance of COVID-19 testing and pandemic in the future

DEDICATION

Goes to my parents James Marachi and Lydia J.Marachi,Siblings and friends who encouraged me and gave me their unwavering support throughout the entire project time.

ACKNOWLEDGMENTS

I would like to extend my gratitude to my advisor Dr. Hansapani Rodrigo for giving me a continuous guidance, support and kindly encouragement throughout the research work. I would also like to appreciate Dr. George Yanev, Dr. Tamer Oraby, Dr. Zhijun Qiao, for accepting to serve as my thesis committee. Acknowledgment also goes to family and all friends who provided moral support throughout this research work.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

1.1 Background

In 2020, COVID-19 became the first pandemic in the world's history to bring the entire world to an abrupt and unexpected halt resulting in major social and economic downturn [22]. Since the first reported case of the disease to date, the novel coronavirus has been able to wreak havoc in literary every corner of the globe and left an ever-growing number of unprecedented fatalities. The normal way of life has been disrupted, and the level of uncertainty about the end of this pandemic continues to manifest to many. Given that the disease is transmitted through fomites, contact, droplets, air, blood, and animal to human, its spread has been rapid within a short period [17]. Area that are either densely populated or remote have equally been affected. Nonetheless, because of the urgency of the matter, numerous and rapid studies have been conducted by researchers worldwide to understand the cause, risk factors, symptoms and mitigation that can be taken against this disease. As various professionals put in more effort, a massive amount of literature and knowledge is swiftly building around this disease.

The advent of corona virus caught many by surprise and this assisted the various to spread faster than it could be contained. The medical professionals rose to the occasion to help fight this deadly pandemic through providing suggestions to the masses to help them prevent the spread of the disease. Some of these suggestions include wearing masks whenever in public, keeping social distance, constantly washing hands and avoiding being in crowded areas [23]. Despite these

recommendations being scientifically proven as effective in combating transmissible diseases, they are not easy to implement and effectively incorporate in the usual way of life. In some cases, people have interpreted some of these recommendations as an infringement of their personal rights. This has resulted in some section of the population neglecting these preventive measures leading to the continued spread of the virus.

The fight against the spread of coronavirus is a collective responsibility that requires socially responsible behavior from people. The health ministry's and other health agencies' advisory on the virus and the doctors' recommendations all over the world insist that the containment of COVID-19 depends on the individual effort that everyone puts to adhere to the guidelines that have been spelled out. When it comes to any infectious diseases, the social behavior of a person plays a massive role in risking one's as well as other people's lives. Failing to take simple precautionary measures such as wearing a mask when in public spaces, keeping social distance and avoiding large crowds whenever necessary can results to exposure to the virus, which can lead to death in the worst-case scenario. In a nutshell, the fight against coronavirus calls for exercising prosocial behavior because the health of other people depends on everyone. According to [4], prosociality is vital in the fight against the pandemic since prosocial individuals are more likely to adhere to COVID-19 guidelines.

Being a transmissible and fatal disease, COVID-19 has been established to have a number of its predisposing risk factors that determine its severity. One of the leading factors that a range of studies has determined is age [36]. Older people of ages above 60 have been found to me at a higher risk of fatalities or sever infection. Other factors in the human body that have been found to increase the severity of the disease include gender, where males are more vulnerable than females, comorbidities such as hypertension and diabetes and race, where black people have been found to suffer more from the disease compared to other races in the United States [13].

In any pandemic whose cure is not yet known, identifying and isolating positive cases is an essential step to combating the spread of the disease. Despite having a low mortality rate of between 0.2% and 9.4% depending on the country based on the Johns Hopkins Coronavirus Resource Center findings, the side effects of the disease are crippling and have been proven to damage people's lives permanently. Research shows a high prevalence of up to 80% of loss of smell in some European countries due to the contraction of COVID-19 [18]. Also, more findings show that approximately 60% of the patients that contracted coronavirus had lost their sense of taste. Despite others regaining their sense of smell and taste, medical practitioners warn that some cases will suffer permanent loss of these senses. To reduce the number of these infections and their side-effects from increasing, identification and isolation of positive cases have proved vital.

1.2 Statement of the Problem

Push for COVID-19 testing has been lauded by both government institutions and health agencies. This inspired the introduction of the mass testing campaign that assists millions to know their status as far as COVID-19 is concerned. Even though large-scale testing of a population can be elusive, it is critical to control a highly transmissible disease such as COVID-19. According to The COVID Tracking Project website, more than half of the US population tested for coronavirus. Even though the number of performed tests seemed to be increasing and on the right track, it was evident that Americans did not unanimously accept the idea. Some individuals are still showing some reluctance to take the test.

Given that prosocial behavior is an essential component in the spread and transmission of coronavirus, it is essential to determine if people are doing enough to keep others and themselves protected from the virus. Taking a COVID-19 test is a prosocial act that is vital in mitigating the spread of COVID-19. Many people flaunt COVID-19 guidelines because they believe that the

virus cannot infect them. Several studies have pointed out that a person can spread the virus to other people if infected, regardless of being symptomatic or asymptomatic. This implies that a person not showing coronavirus symptoms does not mean they are healthy, thus not infectious. Research shows that coronavirus is elusive, and a good number of the infected people tend to have mild symptoms of the virus or be asymptomatic before their infection becomes severe [32]. Research shows that asymptomatic individuals and patients who display mild virus symptoms are the significant propagators of COVID-19 [1]. Therefore, it is vital to conduct a test, which will allow people to know their status and take necessary measures to prevent the spread of the virus.

Testing individuals for coronavirus is an important undertaking in the fight to control the spread of this pandemic. Even as the virus ravages the planet, there has been a huge resistance from the populace when called to take the coronavirus test. Most people who do not want to take the COVID-19 test are afraid of the stigmatization that comes with being found positive and being quarantined, while others are outrightly ignorant. Studies show that one cannot rely on the symptoms of the virus as the methods of ascertaining positivity. Through testing, the asymptomatic patients and those with mild symptoms are made aware of their status, and this can allow them to protect their loved ones and other people. According to [**?**], performing tests for COVID-19 frequently, especially in high-risk areas, can help contain the spread of the virus significantly. Through testing, a broader understanding of the virus can be made, and insights are developed to assist in managing the spread and effects of the virus [26]. Using the COVID-19 longitudinal survey done Understanding America Study, this paper will attempt to unravel some of the factors that affected the willingness of Americans to take the COVID-19 test.

1.3 Statement of the Purpose

As much as they occur once in a long while, pandemics can be relatively common occurrences in the world. As it can be referred through history and the present times, pandemics have a way of crippling human activities resulting in losses of different kinds. COVID-19 is the first pandemic of the 21st century that brought to a halt global life as we know it. In retrospect, it is evident that a lack of preparedness and ignorance orchestrated the spread of this pandemic. Despite the numerous efforts that doctors and health agencies made to inform people on the guidelines that could help contain the pandemic, it was apparent that many were unwilling to do the necessary. Testing emerged to be an essential step that was required to contain the spread of the virus. Nonetheless, many people refused to take the test to determine their status and take necessary preventive measures if needed.

Essentially, testing is instrumental in mitigating the spread of a contagious virus. Through testing, positive cases can be identified, allowing necessary precautions to be taken to slow down the spread of the disease. The mitigation steps taken to control the spread of coronavirus for those who turned up for testing include self-quarantining, where an individual who has been tested positive is provided with a set of rules that have to be adhered to for two weeks. These steps are crucial in safeguarding life besides preventing the spread of the virus. Despite the merits that come with testing, many people did not turn out for testing. This study attempts to model the factors contributing to a person taking responsibility for going for a COVID-19 test. The findings of this study will assist in improving preparedness for pandemics of the future.

CHAPTER II

LITERATURE REVIEW

2.1 COVID-19 Pathogenesis and Its Predisposing Factors

The COVID-19 pandemic has wreaked havoc globally through the widespread morbidity and mortalities caused by it. Since its first reported case towards the end of 2019, the pandemic has spread worldwide, touching all the nations of the earth bringing death and sickness in its wake. The pandemic was quickly established to be caused by a new type of coronavirus known to humankind called SARS-CoV-2 [5]. Just as the SARS-CoV, this new type of coronavirus has been found to use angiotensin-converting enzyme 2 (ACE2) at its main receptor. The ACE2 in the human body is largely expressed in the respiratory epithelium, alveolar, vascular endothelium, macrophages and monocytes [21]. Based on the already done research, SARS-CoV-2 has been established to be mainly transmitted through the respiratory tract [13]. The virus starts multiplying in the upper respiratory tract, and as it continues to manifest in the body, it spreads to all the organs where ACE2 is expressed. This explains why there is some strong clinical deterioration of the kidneys, heart and gastrointestinal tract during the second week after the body becomes infected. The severity of COVID-19 has been established to vary considerably between patients as studies are now showing that there are factors that lead to the variation in the grimness of the disease [5].

In a study done by [6], it was established that individuals that were 60 years and older were approximately 18 times more likely to be at critical or mortal risk of COVID-19 compared to those less than 60 years. [36] found that patients older than 65 years were six times more likely to be at

critical or mortal risk of COVID-19. Despite figures being different from different studies, it is still evident that the age factor is quite significant in coronavirus fatalities. In a study conducted by [28] that consisted of 17.4 million adults in the UK, race was the most significant risk factor. Asians and blacks were the most vulnerable, with age-sex adjusted hazard ratios of 1.95 and 2.17, respectively.

Gender was found to be a significant risk factor as makes were determined in multiple studies to be more susceptible to the severity and fatalities of COVID-19. According to [36], males were 1.76 more likely in critical or mortal risk compared to females. Numerous studies use records on the patients admitted with severe coronavirus and have found that most patients are male. A study was done by [19] on patients in Wuhan found more males were at risk of severe COVID complications. This makes gender a significant risk factor where males are at a higher risk compared to females [31].

Comorbidities such as hypertension, diabetes, respiratory and cardiovascular diseases have also been determined to be significant predisposing factors of coronavirus [36]. [33] iterate that most of the patients who become severely sick due to the COVID-19 have been found to have some underlying illnesses. A good number of them end up succumbing due to these original comorbidities. In a UK study that included more than 500,000 volunteer participants, dementia, diabetes, pneumonia and repression were found to be the most severe comorbidities that could increase the chances of hospitalization to up seven times. These comorbidities make the patients' body vulnerable to the SARS-CoV-2 elusive virus. For instance, diabetes causes impaired phagocytic cell capabilities and an increased level of ACE-2 receptors, which lowers the cell's guard, allowing easy entry of coronavirus [11]. Patients with this set of risk factors have been established to make up the highest number of severe cases and deaths.

2.2 Health Behaviors and Their Effects on Health Outcomes

In the last decade, there has been an increased interest in understanding the social factors and behaviors that affect the health of an individual. [29] affirm that numerous medical studies have come to establish that social determinants have brought a significant health disparity. It is interesting to note that health disparities are influenced considerably by nonmedical factors [35]. Some of these nonmedical factors that have been found to have a significant impact on a person's health include income, health belief, education and political-economic organization of society [29]. The US Department of Health and Human Services, through the Office of Disease Prevention and Health Promotion, established that personal, organizational/institutional, environmental, and policy factors have a considerable effect on the health and health behavior of people [25]. These factors have been found to foster or damage the health of a person.

[29] define health behavior as actions that an individual engages in that affect their health or mortality. These actions can be known or unknown to the doer, and they have health consequences for both the doer and other people. According to [29], some of the actions include physical activity, substance use, diet and health care seeking behaviors, among many others. These health behaviors can be viewed both at a personal level and at a community or group level, and their effects evaluated. Health scholars have come to appreciate that there are factors in the social world that contribute to certain health behaviors in people. [29] agree that demographical factors such as gender, race, and income, among many others, have a significant effect on the health behavior of a person. [14] acknowledges that the social and environmental conditions of an individual or a population cause significant effects on their developmental health.

[34] conducted a study to establish the factors that influenced the health behavior of indigenous Australians due to the significant health disparities that were being found between the

indigenous and non-indigenous Australians. The study was conducted with the assumption that the indigenous people tend to participate in riskier health behavior compared to the non-indigenous people. The study established that history, racism, culture, social networks, social-economic disadvantage and psychological distress significantly and in a complex fashion shape the health behavior of people. According to [34], in their study, low-income earners tend to have inadequate housing that is overcrowded, which negatively affects their health behavior. At the same time, being in an overcrowded household provides intimate social connections, which in turn can have positive impacts such as boosting self-esteem providing some psychological support resulting to a positive effect on an individual's health behavior.

According to a study on the effects of individual factors on health behavior among college students, it was found that students with a greater health concern and better self-perception of their own health status increased their likelihood of adopting positive health behaviors [15]. Also, individuals who majored in medical or health-related fields were found to have better health behaviors than those in other fields. According to [15], critical health literacy is essential in facilitating better health behaviors at the individual level.

The COVID-19 pandemic has been massively devastating largely because of its easiness of spreading and lack of vaccines or proper medication for treating it before turning lethal. Luckily, medical practitioners and researchers have established how the virus spreads making it possible to come up with guidelines that help to contain the spread of the disease. The US public health and various government officials laid out a constellation of new behaviors that people needed to adhere to in an effort to limit the spread of the coronavirus disease [2]. Some of the behaviors that were being advised against included reduction in close physical contact among people and wearing of mask whenever in public spaces. These guidelines have been shared widely as they require strict adherence from the individual to the community level.

[2] conducted a study that was aimed at analyzing social distancing among Americans during the COVID-19 pandemic as a health behavior as a measure to prevent contraction and spread of the disease. The research hypothesized that counties that practiced healthier behaviors before the pandemic would demonstrate greater adherence to the social distancing behavior during the pandemic. The study used data from Cuebig and Streetlight, which was collected from round 15 million mobile devices of individuals who had consented to the study. The analysis showed that the US counties that practices healthier behaviors more before the pandemic exercised a great reduction in movement during the pandemic. Also, [2]established that socioeconomic status was a huge determinant of reduced movements. The study showed that there was higher reduction in movement among counties that were wealthier and more educated compared to others. This implies that income and education play a significant role in adoption of good health behaviors.

COVID-19 has been spreading fast that it could be contained. Nonetheless, just like other transmissible disease, it spread can be limited if contact between people is checked. One of the containment measures that has been applied during this pandemic is self-quarantine due to the high number of cases being recorded each day. The patients that found positive are advised to quarantine themselves if they do not demonstrate severe infection. In that case, the health behavior of an individual is called to question since the choice if voluntary. Moreover, the launching of mass testing all across America was an initiative that was meant to ensure that more Americans are aware of their coronavirus condition and take necessary precautionary measures to protect their loved ones and themselves. This study seeks to find out that factors that influence the choice of an individual to take a coronavirus test.

CHAPTER III

METHODOLOGY

3.1 Study Design and Participants

This study used the descriptive study design where part of the data collected by the Understanding American Study (UAS) in their national longitudinal survey of the attitudes and behaviors around COVID-19 in the USA was utilized. The participants of this survey were randomly sampled using the Address-Based Sampling (ABS) from postal records drawn from across the country. Once selected, the participants were sent a notification letter with a recruitment package, and they were given an option to sign up for either an online or paper survey. The consent of all the participants was obtained to ensure that the survey abides by the ethical rules of research. Eligible participants were required to be adults of 18 years and above, and they were incentivized to respond to the questionnaire. The targeted sample size was 8,900 participants, but only 6,067 of them could complete the questionnaire and this study utilized 440 of the completed responses that had no missing values.

R version 4.1.0 was used to perform data cleaning and analysis on the data in this study. Both descriptive and inferential statistics were computed. For the descriptive analysis, frequencies were obtained as the majority of the variables were categorical. The inferential statistics consisted of bivariate and machine learning (ML) models analysis. Given that nearly all the variables were categorical in nature, the chi-square test was performed as a bivariate analysis for each pair of variables.

3.2 Modeling of the factors that influence willingness to take COVID-19 test

The machine learning techniques that were applied in the development of a model that could predict the willingness of an individual to take coronavirus test include logistic regression, decision tree, gradient boosting and artificial neural network. These techniques were employed using different packages and parameters. The sample data was split into the training and testing dataset, where 80% (n=352) of the data was used in training and the remaining 20% (n=88)was used in the testing dataset. The training data was used to fit the models, which were then tested using the testing dataset. The model development process entailed performing various initializations to each technique. Ten different random initializations were used and the average values of the model were obtained average values of the model performance metrics in the machine learning techniques that were applied. The metrics that were used to evaluate model performance were the accuracy, specificity, sensitivity, Area Under Receiver Operating Characteristics Curve (AUC) and Akaike Information Criterion (AIC) values and the optimum models selected [7].

Logistic regression is type of a statistical regression technique that can be used to analyze data with categorical variables as the response variable [30]. Logistic regression model is used to predict the chance of an event happening and the model is if the form shown below.

Let's consider a model with p number of predictors, $x_1, x_2 \ldots\ x_p$, and a categorical response variable Y, denoted as $\pi = P(Y = 1)$. Let's assume a linear relationship between the predictor variable and the log-odds of the event that $Y = 1$ [8]. The linear relationship between the log odds of the even $Y = 1$ and its predictors can be mathematically written as

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{3.1}$$

where π represent probability of the event $Y = 1$, β_i are the regression coefficients and x_i are the variables.

To make π the subject of the formula, both sides of the model can be exponentiated and the equation below obtained.

$$\pi = \frac{e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}} \tag{3.2}$$

The glm () function in R was used to create the logistic regression models in this study. The fitted model was initialized ten times to obtain the average value of accuracy, specificity, sensitivity, AUC and AIC. A summary of the optimum model was then generated using the summary () function and odds ratio of the significant variables computed together with their 95% confidence intervals.

Classification and regression decision tree model(CART) was the second ML technique to be applied in this study. A decision tree is a supervised machine learning algorithm used in performing both classification and regression analysis. In this study, decision tree was used to perform a classification analysis given that the response variable was categorical. The building of a decision tree generally requires two steps. The process starts with the division of the predictor spaces into J distinct and non-overlapping regions that can be denoted by R_1, R_2,..., R_J [16]. The predictor spaces contain a set of all possible values for X_1, X_2,...X_p. In the next step, the most occurring class for all the observations in the region R_j is obtained and made the prediction for all the observations in that region. Recursive binary splitting is used to grow the classification tree. The binary classification criterion used is the Gini index, which is defined as

$$G = \sum_{k=1}^{K} \hat{\pi}_{mk}(1-\hat{\pi}_{mk}) \tag{3.3}$$

The Gini index is the measure of the total variance across the k classes and $\hat{\pi}_{mk}$ is the proportion of training observations in the mth region that are from the kth class. Gini index is referred to as a measure of node purity as it takes a small value indicating that a node contains largely observation from a single class.

Gradient boosting was the third machine learning classification technique that was used in this study. The gradient boosting classification technique is based on the premise of improving upon a previous weaker model. The method works on the principle of first fitting a model using the training dataset, then improving the model in the subsequent iterations through rectifying the errors of prior models. The improvement of the following models is achieved using a loss function, which is the log-likelihood [10]. The loss function is minimized through the addition of weaker learners using gradient descent.

Using the training dataset, the gradient boosting classification approach starts with fitting a model to predict the dependent variable. The fitting of the gradient boosting classier can be summarized in few steps as follows.

Step 1:The first step of gradient boosting is to build a base model to predict the observations in the traing dataset.

Here,

$$\gamma_i = log(odds)_i = log(\frac{\pi_i}{1-\pi_i}) \tag{3.4}$$

$$F_0(\mathbf{X}) = \arg\min \sum_{i=1}^{n} L(y_i, \gamma_i) \tag{3.5}$$

Where L is the loss function given in equation 3.6 and $\gamma_i, ..., \gamma_n$ are the predicted value. The arg min is used to predict γ value that minimizes the loss function. Here y_i is the observed value and $F_0(x)$ indicate the initial gradient boosting model.

$$L = -\sum_{i=1}^{n} [y_i log(\pi_i) + (1-y_i) log(1-\pi_i)] \tag{3.6}$$

$$L = -\sum_{i=1}^{n} [y_i log(\pi_i) + log(1-\pi_i) - y_i log(1-\pi_i)]$$

$$L = -\sum_{i=1}^{n}[y_i(log(\pi_i) - log(1-\pi_i)) + log(1-\pi_i)]$$

$$L = -\sum_{i=1}^{n}[y_i log(\frac{\pi_i}{1-\pi_i}) + log(1-\pi_i)]$$

$$L = -\sum_{i=1}^{n}[y_i\gamma_i + log(1-\pi_i)]$$

Note that

$$log(1-\pi_i) = log(1 - \frac{e^{(\gamma_i)}}{1+e^{(\gamma_i)}})$$

$$= log(\frac{1+e^{(\gamma_i)} - e^{(\gamma_i)}}{1+e^{\gamma_i}})$$

$$= log(\frac{1}{1+e^{(\gamma_i)}})$$

$$= log(1) - log(1+e^{(\gamma_i)})$$

Since,

log (1) = 0

then,

$$= -log(1+e^{(\gamma_i)})$$

Therefore,

$$L = \sum_{i=1}^{n}[-y_i\gamma_i + log(1+e^{(\gamma_i)}] \tag{3.7}$$

Step 2:The loss function is then differentiated to find the log odds that minimizes the loss function. After differentiation, the minimized loss function is found to be as shown below.

$$\frac{dL}{d\gamma_i} = \frac{d}{d\gamma_i}[\sum_{i=1}^{n}[-y_i\gamma_i + log(1+e^{(\gamma_i)}]] \tag{3.8}$$

$$= \sum_{i=1}^{n}(-y_i + \frac{e^{(\gamma_i)}}{1+e^{(\gamma_i)}})$$

Now, setting

$$\frac{dL}{d\gamma_i} = 0$$

We get,

$$\gamma_i = \frac{e^{(\gamma)_i}}{1+e^{(\gamma)_i}}$$

Step 3:Residuals are then computed to build the next decision tree using all the independent variables. We calculate the residuals based on the following formula for each tree, m which can goes 1 to M.

$$r_{im} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]F(x) = F_{(m-1)}(x) \forall i = 1,2,...,n \tag{3.9}$$

Step 4:Fit a regression tree to the residual values and create terminal regions R_j for $j = 1,2,...,j_m$

Step 5: For each leaf in the new tree, we calculate gamma which is the predicted value. The summation should be only for those records which goes into making that leaf

$$For j = 1,2,...j_m \; compute \; \gamma_{jm} = argmin \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma) \tag{3.10}$$

Since it is possible that one terminal region has many values. After, some modifications we get the generalized formula for gamma.

$$\gamma = \frac{Sum \; of \; residuals}{Sum \; of \; each \; \pi(1-\pi) \; for \; each \; sample \; in \; the \; leaf} \tag{3.11}$$

With this, we get the output values for each leaf in the tree.

Step 6: We update our predictions now using the following iterative formula.

$$Update\ F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{3.12}$$

In the first pass, m =1 and we will substitute $F_0(x)$, the common prediction for all samples i.e. the initial leaf value plus ν, (learning rate) into the output value from the tree we built, previously. The summation is for the cases where a single sample ends up in multiple leaves. We then use this new $F_m(x)$ value to get new predictions for each sample. These new predicted values get a little closer to actual values.It is to be noted that in contrary to one tree in our consideration, gradient boosting builds a lot of trees and M could be as large or more.

Random forest is the fourth machine learning technique that was used in this study. This technique is implemented through bringing together a number of simple decision trees that operate as an ensemble system. Using the various trees in the ensemble, predictions are made based on the class, which is the response variable, and the class with the highest number of predictions is selected to be in the final predicted value [20]. By using a large number of uncorrelated decision trees, the random forest model is able to utilize the power in numbers to produce the best results compared to what one can obtain from an individual tree. This makes the average of all trees an vital feature of the model.

$$RFfi_i = \frac{\sum_{j \varepsilon alltrees} normfi_{ij}}{T} \tag{3.13}$$

where $RFfi_i$ is the importance of the feature, $normfi_{ij}$ is the normalized importance i in jth tree and T is the total number of trees.

Artificial Neural Network was the last machine learning technique to be applied in this study. This technique was inspired by the way human brain functions when processing information. An ANN model can make predictions using both individual variables or a complex interaction among them. They are powerful modeling tools for nonlinear functions and non-additive effects.

There are various types of neural networks, and the feed-forward ANN known as the multi-layer perceptron neural network (MLP) is the one used in this project . Considering an input vector x and target vector D, the MLP used a multivariate nonlinear function to map these input and the target vectors as organized networks consisting of many interconnected layers [27]. Each of these layers comprises a set of artificial neurological connections that do not have any feedback loops. The MLP technique uses both inputs and outputs for training purposes. The architecture of an MLP with d inputs, M hidden layers and K output nodes is shown in the figure 3.1 below.

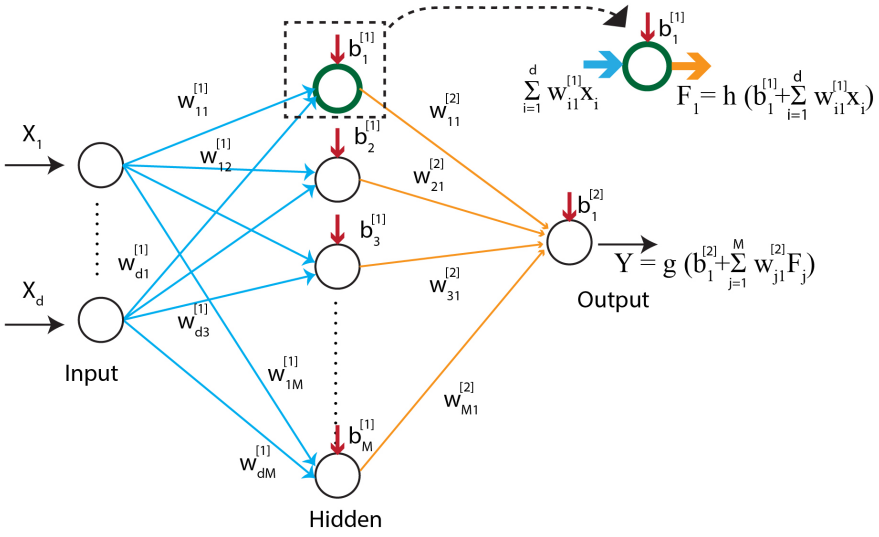

Figure 3.1: Architecture of Multi-layer Perceptron Neural Network

The analytical function of an ANN model is developed through the following process. The input of the j^{th} hidden unit, say, $a_j^{[1]}$ is formed through a weighted linear combination of d input values and their associated hidden unit bias which is represented by the equation 3.14 below

$$a_j^{[1]} = \sum_{i=1}^{d} w_{ji}^{[1]} x_i + b_j^{[1]} \tag{3.14}$$

Where, $w_{ji}^{[1]} x_i$ is the associated weight of the input i and the hidden unit j, and b_j is the bias of the j^{th} hidden unit.A non linear differentiable activation function $h(.)$ is applied on equation 3.14

to get the activation of the hidden unit j.

$$z_j = h(a_j^{[1]}), j = 1,2,3,...M \tag{3.15}$$

The hyperbolic tangent activation function below is used due to its faster convergence.

$$h(a_j^{[1]}) = tanh(a_j^{[1]}) \tag{3.16}$$

A weighted combination of z_j and the corresponding bias associated with each output node $b_j^{[2]}$ form the input $a_k^{[2]}$ to each output node

$$a_k^{[2]} = \sum_{j=1}^{M} w_{kj}^{[2]} z_j + b_k^{[2]}, k = 1,2,...K \tag{3.17}$$

A nonlinear transformation $g(.)$ is applied to obtain the final outcome as shown in equation 3.18 below

$$y_k(\mathbf{x},\mathbf{w}) = y_k = g(a_k^{[2]}) \tag{3.18}$$

Given the target variable is binary, the activation function g(.) is the logistic sigmoid function of the form.

$$y_k(\mathbf{x},\mathbf{w}) = \frac{1}{1+exp(-a_k^{[2]})} \tag{3.19}$$

The final activation form of the MLP for the k^{th} output node can be written as:

$$y_k(\mathbf{x},\mathbf{w}) = g(a_k^{[2]}) = g\Big(\sum_{j=1}^{M} w_{kj}^{[2]} h\Big(\sum_{i=1}^{d} w_{ji}^{[1]} x_i + b_j^{[1]}\Big) + b_k^{[2]}\Big). \tag{3.20}$$

3.2.1 Feature Importance

Aside from logistic regression approach, the other machine learning techniques used in this study tend to have their complexities when it comes to understanding them and interpreting their results. However, due to the great merits that this techniques bring based on their performance metrics, it is always important to consider them. In this study, the model-agnostic post-hoc framework SHAP has been used to help with the interpretation of this complex algorithms through estimating the importance of features [24]. The SHAP framework is made up of model prediction as a sum of SHAP values of each feature as shown below.

$$\zeta(x) = \phi_0(\zeta, x) + \sum_{i=1}^{M} \phi_1(\zeta, x) \tag{3.21}$$

Where, $\phi_0(\zeta, x)$ is the model bias and M is the number of features. The SHAP values are computed using the best model, ζ_{opt}, with the optimum AUC. The ranking of the features in the best model, ζ_{opt}, was obtained by calculating the mean SHAP value magnitude over all the instances as shown in the equation below.

$$A_j = \sum_{i=1}^{N} |\phi_j(\zeta_{opt}, x)| \tag{3.22}$$

Where, A_j is the attribution of the j^{th} feature.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 Results

4.1.1 Descriptive analysis results

From the sample of 440 respondents, 57.95% of the individuals were female while 42.05% were male, which shows a good representation of both genders. 54.32% of the sampled respondents stated that they were married, while 45.68% claimed not to be in any form of marriage.

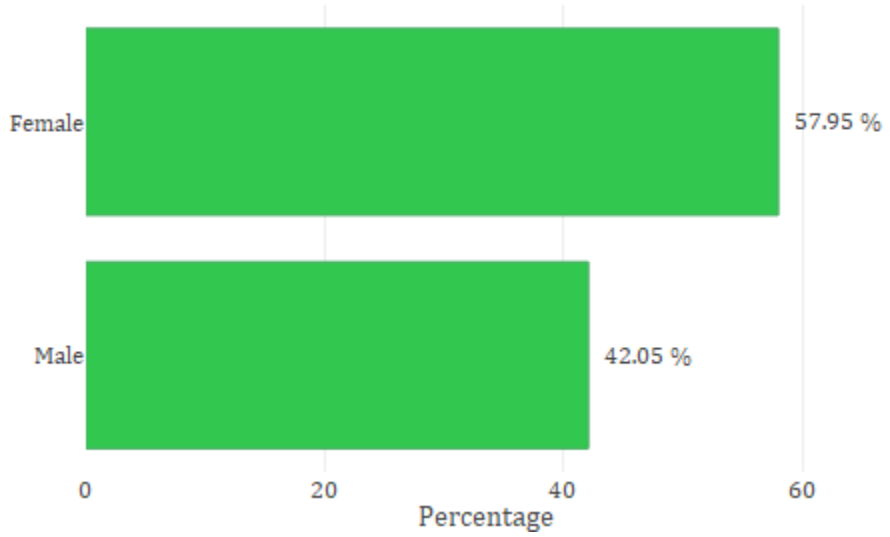

Figure 4.1: Gender of the Respondents

The sample can be considered to contain mostly middle-aged individuals, but with a wide variation in their ages ($\bar{X}$=52.54, SD=16.02). The analysis showed that majority of the respondents (75.23%) were white, while the other races shared the remaining portion.

Most of the respondents (80.68%) of the respondents were found to have some sort of tertiary level education, and a good number of them (16.95%) had secondary level education. Majority of the respondents (62.95%) of the respondents claim to be working, and the remaining portion (37.05%) claimed not to have any form of employment.

Table 4.1: Education Level of the Respondent.

Education Level	**n**	**%**
Primary Level Education	12	2.73
Secondary Level Education	73	16.59
Tertiary Level Education	355	80.68

According to the analysis, most of the respondents (51.59%) earned an income of between $30,000 and $99,000. A good number of them (27.05%) earned more than $99,000 annually, while a small portion (21.36%) earned less than $30,000. These findings show that most of the respondents were earning more than the average income in the US.

Table 4.2: Household Income of the Respondents

Income Level	**n**	**%**
$ 100,000 and above	119	27.05
Below $30,000	94	21.36
Between $ 30,000 and $ 99,000	227	51.59

Race of the respondents was a crucial demographical factor that had to be included in the study. From the already existing literature, it can be deduced that race is an important issue when it

comes to the coronavirus spread and treatment. The output below provides the results of the racial distribution of the respondents.

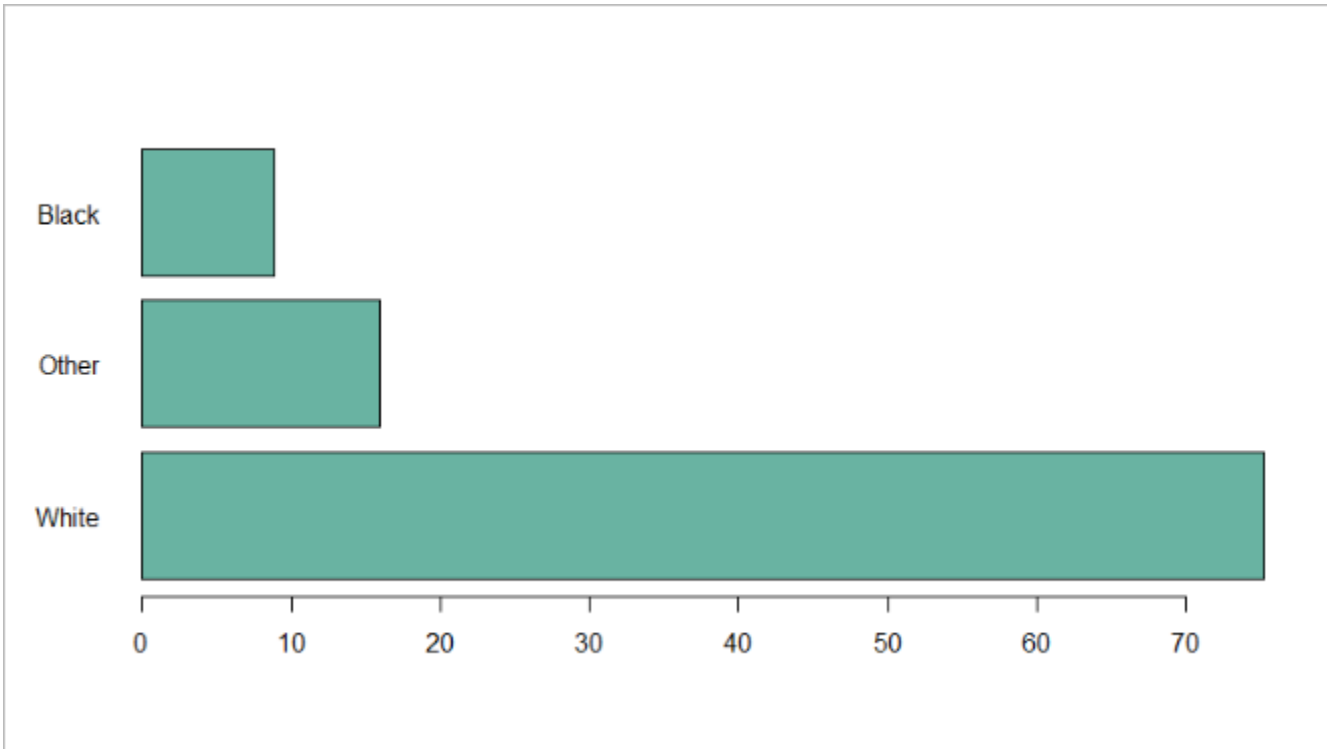

Figure 4.2: Race of the Respondents

Based on the output in figure 4.2, it can be seen that the majority of the respondents were whites, constituting more than 75% of the sample. Less than 10% of the respondents were black and the rest were other races. Even though the sample was predominantly white,followed by other categories it shows how the America population is distributed.

The study also wanted to establish the immigration status of the respondents. Immigration status is essential especially when it comes to health care accesses. A bar graph below was generated to show the immigration status of the respondents.

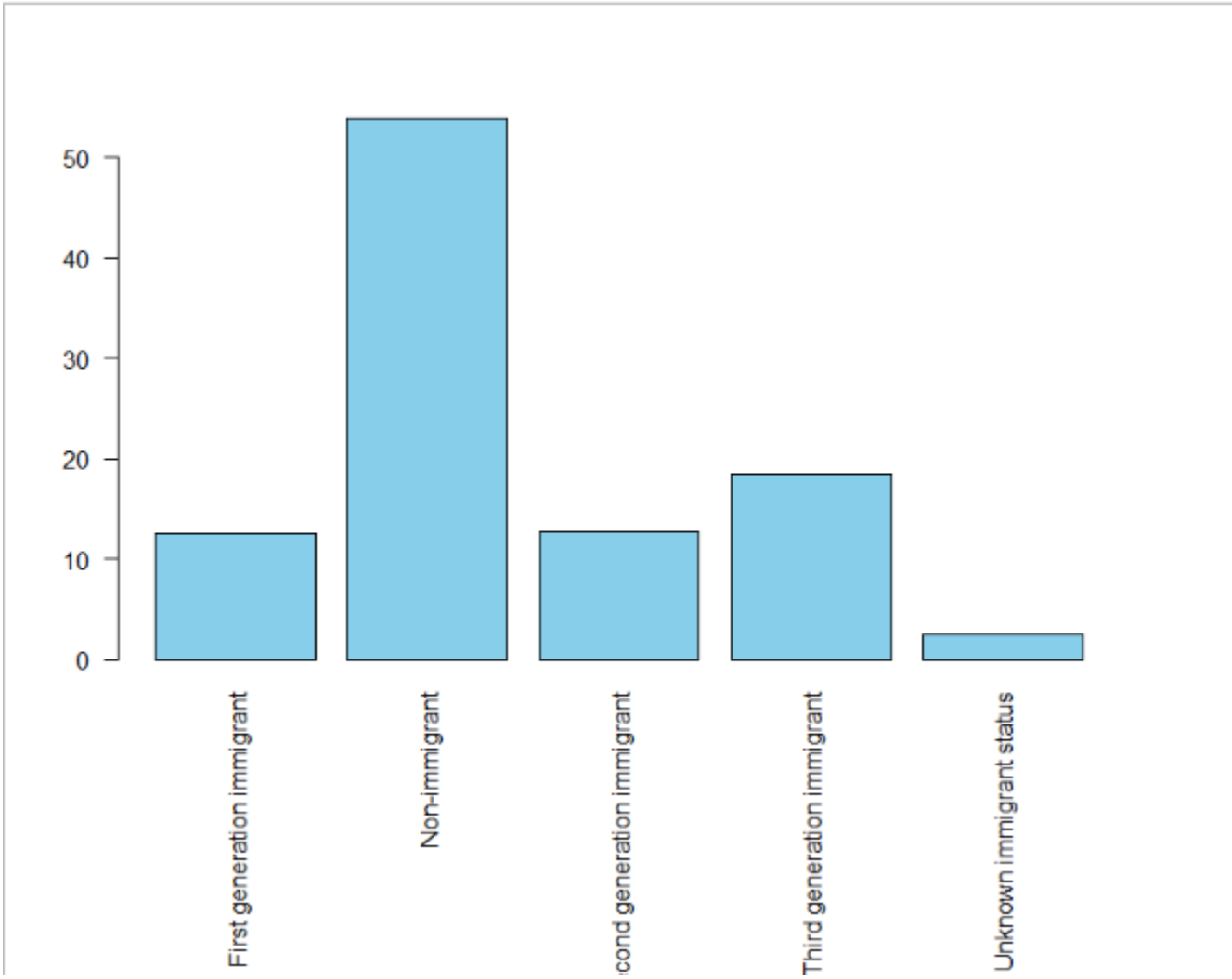

Figure 4.3: Immigration Status of the Respondents

As it can be seen in figure 4.3 above, majority of the respondents (53.8%) were none immigrant. The sample consisted of 12.5%, 12.7% and 18.4% first, second and third generation immigrants. A tiny portion of the sample (2.5%) stated to have unknown immigration status. During the pandemic, the economy received a major blow as business could not be conducted as usual. Due to this, many companies and business institutions were forced to lay of workers to ensure that they could remain afloat. This study wanted to establish the number of individuals that still had their jobs and those who were unfortunately unemployed and the results are as shown in the figure 4.4

pie chart.

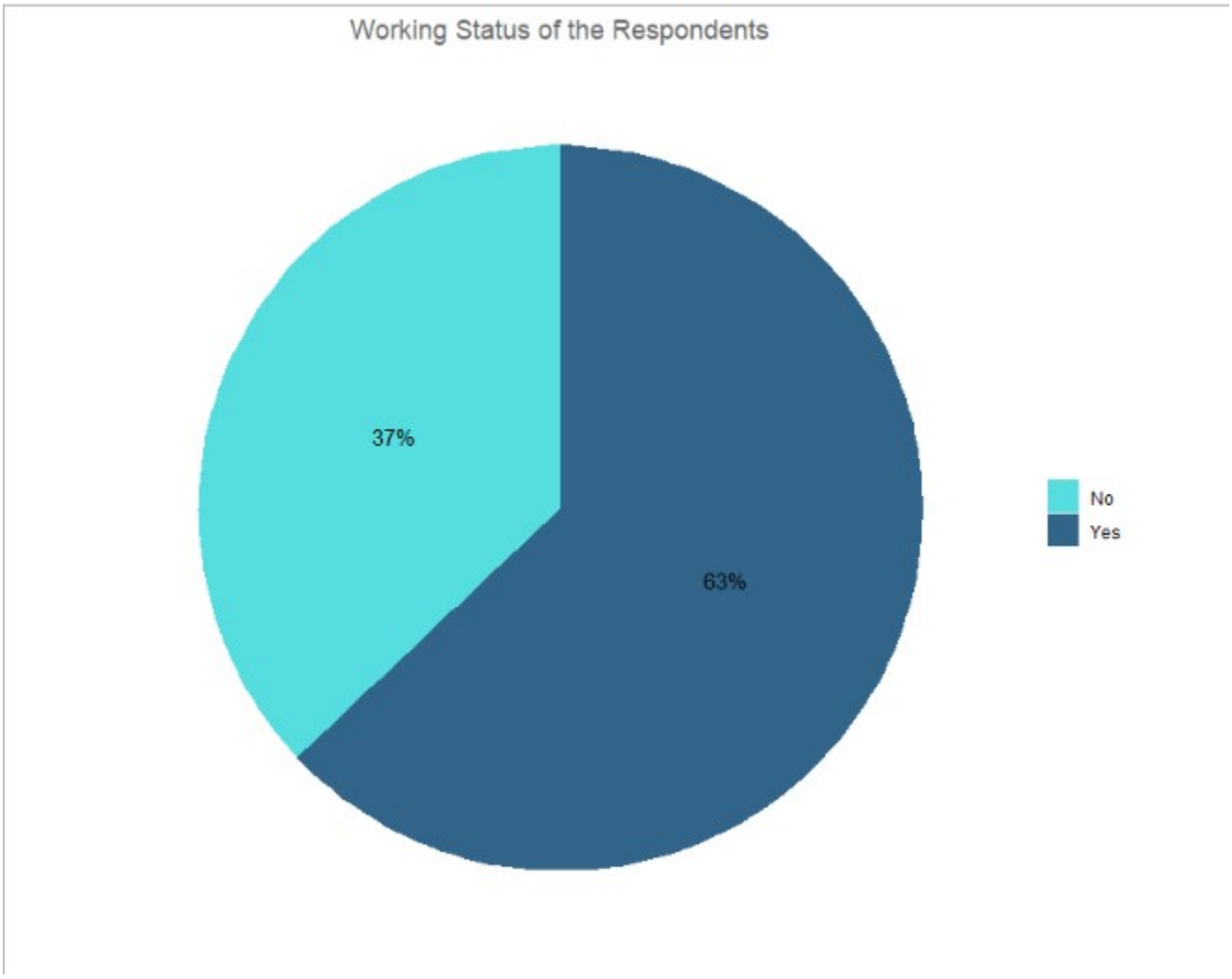

Figure 4.4: Working Status of the Respondents

From figure 4.4 above, it can be seen that majority of the respondents (65%) stated that they still have their jobs and are working. Also, the finding shows that 37% of the respondents were not working. As much as most of the respondents claimed that they were still working at their respective jobs, it was shocking to see such a high number of respondents claiming not to be working.

Nearly, all the respondents (92.05%) stated to have health insurance cover. Such a high

percentage of insured persons was expected given that most of them earned a descent income. On the issue of the source of information about coronavirus, 20% claimed to obtain their information from social media, 47.95% from media outlets, 58.83% from health entities, 24.55% from their physician and 47.95% from family, friends and colleagues. These results indicate that health entities, media outlets, family friends and colleagues are the most utilized sources of information on COVID-19 by most Americans.

4.1.2 Bivariate Analysis

This study sort to find out factors that could contribute to an individual taking a COVID-19 test. Given that numerous studies have already been done on this pandemic, much has been revealed about the perception about the disease and the steps that individuals are taking to prevent themselves from contracting it. This study attempts to understand the commitment of the American populace towards taking corona virus test as an initial step in their protection. The research starts by analyzing the association between the characteristics of the sampled individuals and their COVID-19 test status.

Besides providing the breakdown of the demographic characteristics of the sample and how they relate to the state of being tested for corona, table 1 above shows the chi-square results of each characteristic based on whether an individual has undergone a corona test or not. As it can be seen, not all demographic characteristics had some significant association with the COVID-19 testing status. Gender (0.030), marital status (0.002), race (0.000) and income (0.032) were found to be significantly associated to taking a COVID-19 test at 0.05 level of significance, while education (0.784) and working status (0.393) were insignificant.

As it can be seen in table 4.3, for all the categories in gender, majority of the respondents claimed not to have taken part in getting tested for COVID-19. This can be seen in other demographic characteristics such as income, education, marital and working status. Interestingly, not all categories of race had the same pattern. From the table 4.3 above, it can be seen that out of the

Table 4.3: Demographic characteristics of the respondents and their association to taking COVID-19 test.

Characteristic		**Total**	**Not Tested**	**Tested**	**p-value**
Gender, n (%)					0.030
	Male	185 (42.05%)	149 (33.86%)	36 (8.18%)	
	Female	255 (57.95%)	181 (41.14%)	74 (16.82%)	
Marital status, n (%)					0.002
	Married	201 (45.68%)	194 (44.09%)	45 (10.23%)	
	Not Married	239 (54.32%)	136 (30.91%)	65 (14.77%)	
Education, n (%)					0.784
	Primary Level	12 (2.73%)	10 (2.27%)	2 (0.45%)	
	Secondary Level	73 (16.59%)	54 (12.27%)	19 (4.32%)	
	Tertiary Level	355 (80.68%)	266 (60.45%)	89 (20.23%)	
Race, n (%)					0.000
	White	331 (75.23%)	258 (58.64%)	73 (16.59%)	
	Black	39 (8.86%)	18 (4.09%)	21 (4.77%)	
	Other	70 (15.91%)	54 (12.27%)	16 (3.64%)	
Working, n (%)					0.393
	Yes	277 (62.95%)	212 (48.18%)	65 (10.23%)	
	No	163 (37.06%)	118 (26.82%)	45 (10.23%)	
Income, n (%)					0.032
	Below $30,000	94 (21.36%)	64 (14.55%)	30 (6.82%)	
	$30,000 - $99,000	227 (51.59%)	167 (37.95%)	60 (13.64%)	
	$100,000 and above	119 (27.05%)	99 (22.50%)	20 (4.55%)	
Age, Mean (SD)		51.70 (15.73)	52.78 (15.46)	48.47 (16.14)	0.723

whites the majority of them are not tested while, amongst the black respondents there was a slight increment in the percentage of those who get tested corona test. Nonetheless, the results apparently show that the rate of testing for corona is low amongst the various demographics in the US.

Table 4.4: Source of corona virus news and its association with COVID-19 testing.

Variable		**Total**	**Not Tested**	**Tested**	**p-value**
News from social media, n (%)					0.999
	Yes	88 (22.75%)	66 (15.00%)	22 (5.00%)	
	No	352 (77.25%)	264 (60.00%)	88 (20.00%)	
News from Media Outlets, n (%)					0.295
	Yes	211 (47.95%)	153 (34.77%)	58 (13.18%)	
	No	229 (52.05%)	177 (40.23%)	52 (11.82%)	
News from Health Entity, n (%)					0.693
	Yes	269 (61.14%)	204 (46.36%)	65 (14.77%)	
	No	171 (38.86%)	126 (28.64%)	45 (10.23%)	
News from family/friends, n (%)					0.295
	Yes	211 (47.95%)	153 (34.77%)	58 (13.18%)	
	No	229 (52.05%)	177 (40.23%)	52 (11.82%)	
News from Physician, n (%)					0.000
	Yes	108 (24.55%)	64 (14.55%)	44 (10.00%)	
	No	332 (75.45%)	266 (60.45%)	66 (15.00%)	

In table 4.4 above, the various sources of news on corona virus have been tabulated against the COVID-19 test status. During the pandemic, it became apparent that there was some conflicting information about the disease. This confusion led to different individuals having varying ideas about the virus, some beneficial, some damaging. Therefore, it became important to investigate the effect of the sources on information to the act of an individual taking the corona virus test. As it can be seen from table 4.4 above, obtaining news from the physician was the only significant information source that influenced testing for corona virus among the respondents. According to the findings, individuals that obtained their information about the pandemic from their physicians are more likely to take a COVID-19 test compared to those who do not. Also, the findings show

that social media, media outlets, health entities, family, friends and colleagues do not significantly influence the act of an individual to take a corona virus test.

Table 4.5: Association between health insurance and medical condition and taking COVID-19 test.

Variable		**Total**	**Not Tested**	**Tested**	**p-value**
Have health insurance, n (%)					0.723
	Yes	405 (92.05%)	303 (68.86%)	102 (23.18%)	
	No	29 (6.59%)	23 (5.23%)	6 (1.36%)	
	Unsure	6 (1.36%)	4 (0.91%)	2 (0.45%)	
Medical condition, n (%)					0.374
	Yes	260 (56.82%)	183 (41.59%)	67 (15.23%)	
	No	190 (43.18%)	147 (33.41%)	43 (15.23%)	

Majority of the respondent (92.05%) stated that they had some insurance cover, while a small portion(1.36%) claimed that were not certain about their health cover. The bivariate analysis indicated that having a health insurance was not a factor that influenced a person's choice to take a COVID-19 test. A good number of the interviewed persons (56.82%) attested to having one or more health conditions. The study wanted to establish if a person suffering from a health condition would be more likely to take a corona virus test since the reports of the fatalities and complication cases of COVID-19 patients indicated that patients that suffer more are the ones who had underlying conditions such as hypertension, diabetes and obesity. However, the bivariate analysis pointed out that having a medical condition did not influence the respondents to take COVID-19 test.

The study sought to find out the emotions of the respondents given that the pandemic had drastically changed people's lives globally. Despite the hardships that came with the pandemic, the study showed that most of the respondents were in control of their emotions. Majority of the respondents indicated that the pandemic had not negative emotional impact on them. However, a good number of the respondents stated that they had some emotional problems. The bivariate analy-

Table 4.6: Association of Mental health state and taking COVID-19 test.

Variable		Total	Not Tested	Tested	p-value
Feeling nervous, n (%)					0.043
	Not at all	281 (63.86%)	233 (50.68%)	58 (13.18%)	
	Several days	118 (26.82%)	79 (17.95%)	39 (8.86%)	
	More than half the days	18 (4.09%)	13 (2.95%)	5 (1.14%)	
	Nearly every day	23 (5.23%)	15 (2.95%)	8 (1.82%)	
Cannot stop worrying, n (%)					0.008
	Not at all	322 (73.18%)	255 (57.95%)	67 (15.23%)	
	Several days	82 (18.64%)	51 (11.59%)	31 (7.05%)	
	More than half the days	13 (2.95%)	8 (1.82%)	5 (1.14%)	
	Nearly every day	23 (5.23%)	15 (2.95%)	7 (1.59%)	
Hopeless and Depressed, n (%)					0.017
	Not at all	309 (70.23%)	242 (55.00%)	67 (15.23%)	
	Several days	88 (20.00%)	56 (12.73%)	32 (7.27%)	
	More than half the days	24 (5.45%)	20 (4.55%)	4 (0.91%)	
	Nearly every day	19 (4.32%)	12 (2.73%)	7 (1.59%)	
Little pleasure, n (%)					0.543
	Not at all	327 (74.32%)	251 (57.05%)	76 (17.27%)	
	Several days	75 (17.05%)	52 (11.82%)	23 (5.23%)	
	More than half the days	17 (3.86%)	12 (2.73%)	5 (1.14%)	
	Nearly every day	21 (4.77%)	15 (3.41%)	6 (1.36%)	

sis showed that the feeling of nervousness, constant worrying, hopelessness and depression were significantly related to a person taking a corona virus test. These results indicate that individuals whose emotions were negatively impacted, either weakly or strongly, by the pandemic were more likely to take a corona virus test compared to those who were not.

The survey sort to find out if the pandemic had affected the respondents' ability of having things under control in their live. In that regard, several questions were asked about how they felt about having the ability to control things in their lives. The analysis, as shown in table 4.7 above, showed that the ability of individuals to feel control of things in their lives does not have an effect on whether they will be more willing to take the COVID-19 test.

Table 4.7: Association of Mental health state and taking COVID-19 test.

Variable		Total	Not Tested	Tested	p-value
Feeling nervous, n (%)					0.043
	Not at all	281 (63.86%)	233 (50.68%)	58 (13.18%)	
	Several days	118 (26.82%)	79 (17.95%)	39 (8.86%)	
	More than half the days	18 (4.09%)	13 (2.95%)	5 (1.14%)	
	Nearly every day	23 (5.23%)	15 (2.95%)	8 (1.82%)	
Cannot stop worrying, n (%)					0.008
	Not at all	322 (73.18%)	255 (57.95%)	67 (15.23%)	
	Several days	82 (18.64%)	51 (11.59%)	31 (7.05%)	
	More than half the days	13 (2.95%)	8 (1.82%)	5 (1.14%)	
	Nearly every day	23 (5.23%)	15 (2.95%)	7 (1.59%)	
Hopeless and Depressed, n (%)					0.017
	Not at all	309 (70.23%)	242 (55.00%)	67 (15.23%)	
	Several days	88 (20.00%)	56 (12.73%)	32 (7.27%)	
	More than half the days	24 (5.45%)	20 (4.55%)	4 (0.91%)	
	Nearly every day	19 (4.32%)	12 (2.73%)	7 (1.59%)	
Little pleasure, n (%)					0.5433
	Not at all	327 (74.32%)	251 (57.05%)	76 (17.27%)	
	Several days	75 (17.05%)	52 (11.82%)	23 (5.23%)	
	More than half the days	17 (3.86%)	12 (2.73%)	5 (1.14%)	
	Nearly every day	21 (4.77%)	15 (3.41%)	6 (1.36%)	

4.1.3 Model Results

The construction of the models was done using several machine learning techniques namely Logistic regression,CART,gradient boosting, Random Forest and Artificial Neural Network.

4.1.4 Logistic Regression Model

The first technique to share in this section is the bivariate logistic regression model. All 24 variables in the data were used to fit the logistic regression in this research where 23 of them were independent variables.The optimum regression model was selected based on AIC value of 356.67 with a residual deviance of 326.67. The fitted regression model was found to have an AUC value of 0.6074 for the testing data set with a sensitivity value of 78.26% and specificity value of 28.57%. Table 4.8 represent the estimated coefficients of fitted logistic regression model.

In the above table 4.8, it can be seen that only 5 variables were significant predictors of the

Table 4.8: Association of Mental health state and taking COVID-19 test.

Variable	Categories	Estimate	Std. Error	z value	p-value
(Intercept		1.7576	1.0452	1.681	0.0927
Age		-0.0304	0.0097	-3.148	0.0016
Race					
	Black	1.1821	0.4457	2.652	0.0080
	Other	-1.2008	0.37338	-0.538	0.5908
Your physician					
	Yes	1.4271	0.3458	4.127	0.000
Unable to control important things in life					
	Never	-1.2888	0.4228	-3.048	0.0023
	Sometimes	-1.1953	0.5565	-2.148	0.0317

willingness of an individual to take a COVID-19 test. A unit increase in age significantly (P <0.05) reduced the odds of taking a coronavirus test by 3%. Younger individuals appeared to be more likely to take a coronavirus test compared to older people. Black people were found to significantly (P <0.05) have a higher odds of taking a coronavirus test compared to the white people. According to the model, a black person was 3.26 times greater odds of getting tested compared to white person

The source of news and information about coronavirus also showed to have some significant effects taking the COVID test. Obtaining news from social media seemed to significantly (P <0.05) reduce the odds of an individual taking a COVID-19 test. Based on the model results, a person who obtains their information and news about coronavirus through social media was 58.7% less odds to take a coronavirus test compared to one who does not rely on social media. For individuals who obtain information and news about COVID-19 from their physician, there was a significantly (P < 0.05) higher chances that they would take the test. Those who relied on news and information from their physicians about coronavirus were 4.17 higher odds of taking COVID-19 test compared to those who do not. Lastly, the results of the logistic regression model show that people who were mentally feeling that they have never been able to control their life during the pandemic were 72% of lower odds of taking coronavirus test.

4.1.5 Decision Tree Model

The next model to be fitted was the decision tree model. The model performance metrics of the decision tree model emerged to be better compared to the logistic regression model. To obtain average values for the different model performance metrics, 10 random models were run with different initializations. The fitted decision tree model had an average AUC value of 0.6171 with a sensitivity value of 96.97% and a specificity value of 22.73%. The decision tree attempts to break down the various probabilities that comes with being in a certain category. As it can be seen in the root node, an individual picked randomly from a sample of non-tested persons is 75% more likely to be white.

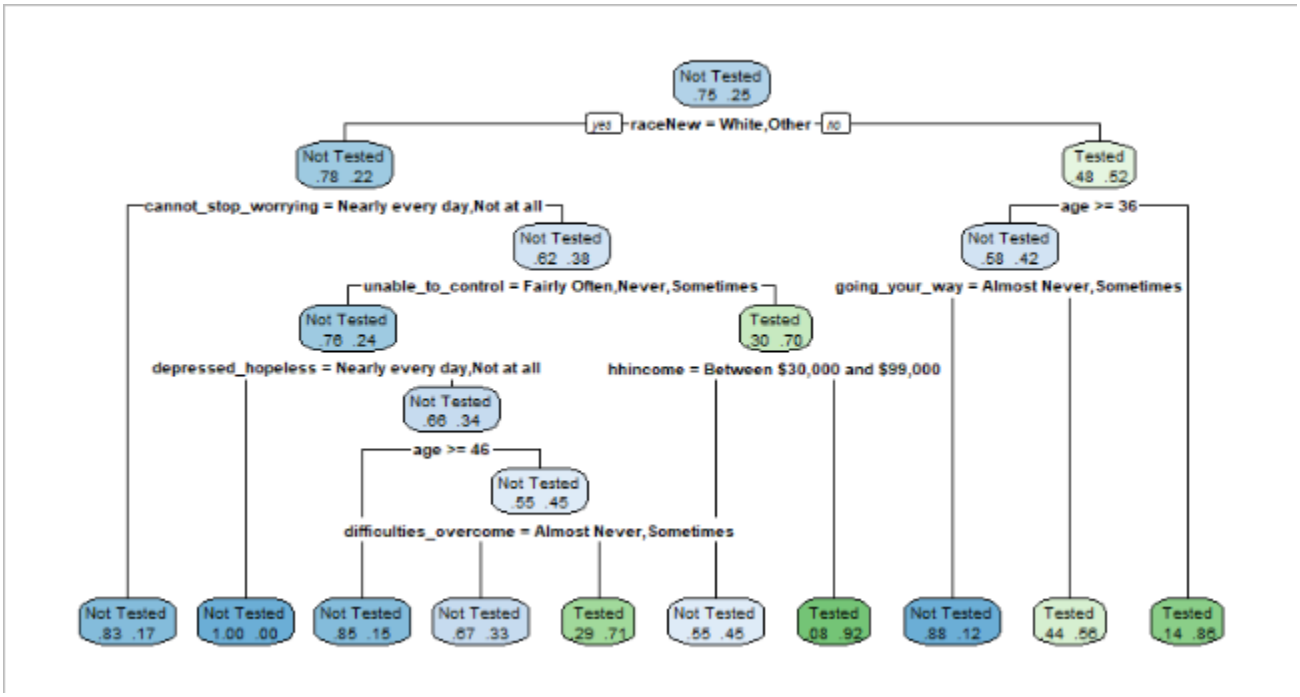

Figure 4.5: Decision tree model

From the above decision tree, it can be seen from the root node that 75% of the respondents were not tested, and 25% of them were tested. The variable being utilized by the model to split and grow the model is the coronavirus test status. The left split of the root node is a branch that is

based on the respondents that were white, while the right branch is based on the other races. As seen in the first decision node in the left branch of the root node, 78% of the white respondents did not take the COVID-19 test, while 22% of them did take the test. In the right branch of the root node, it can be seen that 48% of the individuals from the other races had taken the coronavirus test, while 52% of them had not. Among the respondents that belonged to other races, the majority of the individuals 36 years or older (36%) had not taken the test, while a small portion (14%) did.

Based on figure 4.6 below on the variable importance, the feeling of being unable to control life during the pandemic became a significantly strong impetus that made people go to take a coronavirus test. This was followed closely by the feeling of difficulties to overcome challenges faced during this pandemic. Age was obtained to be the third most significant predictor of whether a person takes a covid test or not. Sources of news on coronavirus and marital status are among the variables that were found to be least significant in the model. The variable importance plot generally showed that the mental state of the individuals played a significant role in determining their action to take a COVID-19 test. Sources of information about coronavirus seemed not to strongly influence the decisions of the respondents to take the coronavirus test.

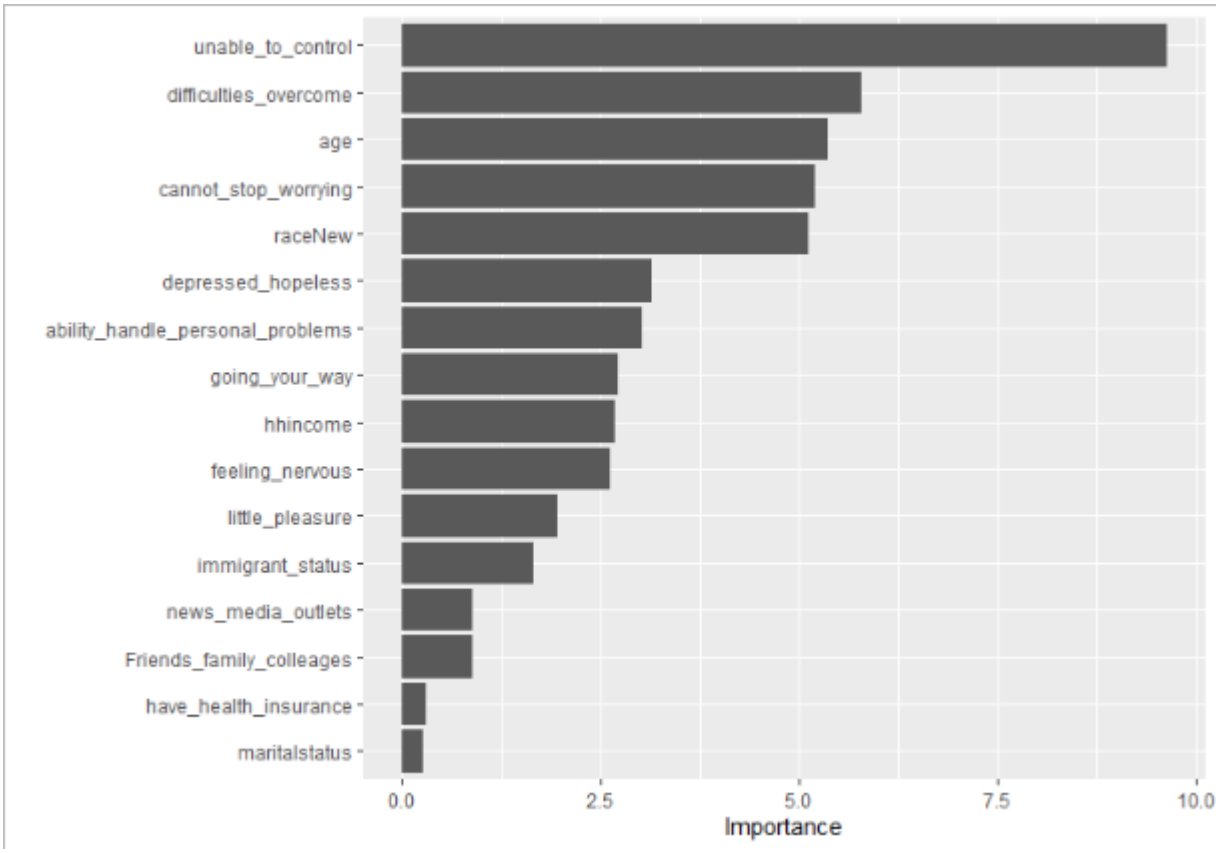

Figure 4.6: Decision tree model variable importance

4.1.6 Gradient Boosting Model

The gradient boosting technique was also used to model the factors that can influence the willingness of an individual to get tested of corona. The fitted model had an accuracy level of 72.73% , which is relatively good, with a specificity and sensitivity value of 75% and 25% respectively. Also, the model was established to have an average AUC value of 0.6102, which is better than a model fitted by chance. Further, an analysis into the variable significance was made and the output in figure 4.7 below was generated.

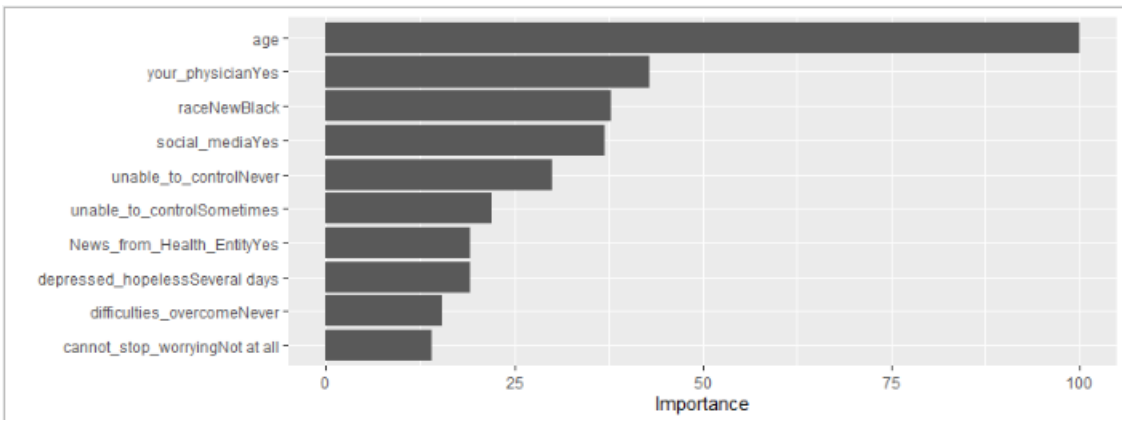

Figure 4.7: Xgboost model

As it can be seen in the above figure 4.7 output, age was established as the most important feature in predicting the willingness of a person to take a coronavirus test. This was followed by obtaining news from physicians. Race and use of social media to obtain information were also found to be highly influential. In this model, it can be seen that mental health did not play a very significant role compared to what was observed in the decision tree model.

4.1.7 Artificial Neural Network (ANN)

Another ML technique to be applied in the analysis was ANN. The fitting of the ANN model encompassed running of multiply running of the model under different set-seed value to obtained the value that produced the best model. The ANN model in this case was fitted using the neuralnet library and it had 3 hidden layers. After running the model multiple times, the ten best models were selected. On average, the model had an accuracy value of 46.21% (±14.38%), a sensitivity value of 50.76% (±23.39%) a specificity value of 44.54% (±19.96%) and an AUC value of 57.35% (±3.87%).

4.1.8 Random Forest Model

The final Machine learning technique to be applied on the data was the random forest technique. Just like the previous models, the random forest model was also run several times to

generate average of the vital performance metrics. After running the model several times using different set seed values an average accuracy value of 76.34%(SD:2.13%) was obtained. Also the mean sensitivity value of 82.61%(SD:0.08%) and specificity value of 13.12%(SD:0.05%) was obtained. Lastly, the model produced an average AUC value of 78.15%(SD:1.05. On plotting the model the graph below was generated.

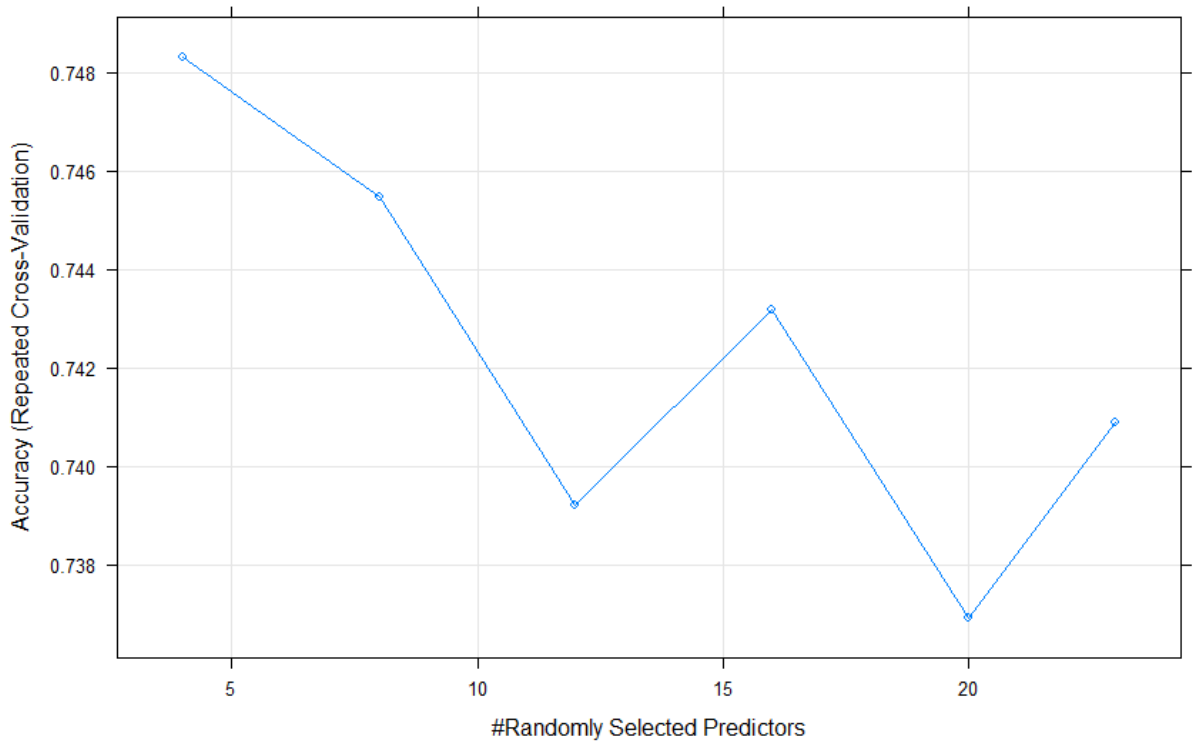

Figure 4.8: Accuracy levels of the random forest model based on different combinations of predictors

According to figure 4.8, it can be seen that the model has a higher accuracy when only 3 predictors are used. The model performance based on the accuracy measure decreases as extra predictors are included in the model. The model performance is at its lowest when only 20 predictors are used. To establish the importance of the different predictors in the model, a feature importance plot below was generated using Shap Analysis technique.

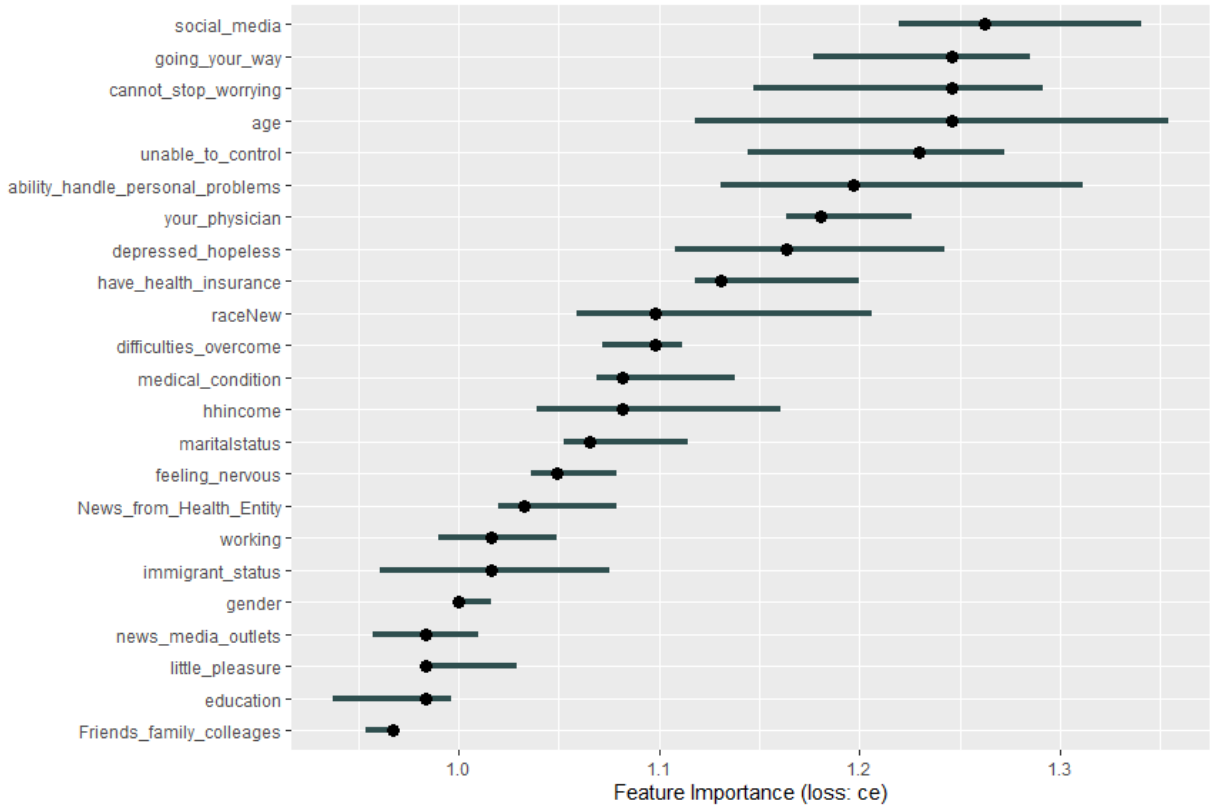

Figure 4.9: Feature Importance of the Random Forest Model

4.1.9 Best Model

Based on the outputs of the above model, the table of model performance evaluation was formed as shown below.

Table 4.9: Model Performance Evaluation

Classification Model	**Accuracy**%	**Sensitivity%** (True Positive Rate)	**Specificity%** (True Negative Rate)	**Area Under the ROC Curve** (AUC)
C5.0 Decision Tree Model	72.39 (±2.89)	90.45 (±6.15)	18.18 (±11.33)	60.91 (±4.51)
Gradient Boosting	72.73 (±14.38)	75.00 (±23.39)	25.00 (±19.96)	61.02 (±3.87)
Logistic Regression	52.95 (±4.71)	83.22 (±4.5)	31.00 (±3.25)	62.60 (±4.13)
Artificial Neural Network	49.21 (±14.38)	50.76 (±23.39)	44.54 (±19.96)	57.35 (±3.87)
Random Forest	76.34 (±2.13)	82.61 (±0.08)	13.12 (±0.05)	78.15 (±1.05)

4.2 Discussion

4.2.1 Key Findings and Their Interpretations

This study utilized data collected from the Understanding America Study COVID-19 survey conducted on April 2021 to investigate some of the factors that influence a person's willingness and responsibility to take a COVID-19 test. The wake of the corona virus pandemic has brought serious responsibility to people given that individual responsibility plays a significant role in controlling its spread. As it can be seen in the results section, the fitted model and the bivariate analysis have shown some of the factors that significantly predict person's willingness to take a coronavirus test. Some of the factors that were investigated in this research include the demographical factors, source of information about coronavirus, medical and psychological issues.

As it can be seen from this pandemic, it is vital for the world to live in preparation for another outbreak based on the lack of preparedness that was demonstrated by the government and the people

in the United States and the world as a whole. As studies are piling up on this particular disease, the understanding of the virus continues to grow. Compared to January 2020, the information of how the virus spreads, the predisposing factors that increase the severity of the disease on an individual and the guidelines of protection are clearer and understandable by the general population. This pandemic has shown that the general public plays a significant role in mitigating the spread of a disease.

At it was clearly evident in 2020 and some part of 2021, many governments resulted to taking severe measures to control the masses during the COVID-19 pandemic because of the spread, morbidities and fatalities that were associated with the disease [4]. In as much as these regulations and restrictions were quite essential in controlling the spread of the disease, they had some massive ramifications on the social aspects of people's lives. These restrictions sparked a backlash as millions of Americans were feeling their rights were being infringed. The backlash was quite serious necessitating an alternative approach that could allow people to go into the public and socialize in a more cautious manner. Relaxing of the initial harsh restrictions that were imposed by the national and the local governments called for trust in people hoping that they would willingly consider the social impact of their health behavior [12].

In a study done by [4], it was found that a good number of Americans do consider the social impact of their behavior during the coronavirus pandemic. It was found that people do assess the effects of their action and are incentivized not to put others at risk, especially if the risk is quite significant. Such prosocial behavior can have a significant effect on the health behavior of an individual assisting them to make the right choices to control the spread of coronavirus. In this study, this finding is brought to question as it is evident that most people did not engage in testing for COVID-19. Only 25% of the respondents saw the need to take a COVID-19 test indicating a small number of prosocial people. As an essential health behavior during this pandemic, taking a

COVID-19 test is vital in informing a person if they are a risk to other people or not. Prosocial people are more likely to participate in actions that help in containing the spread and testing is a crucial step.

The social world has been found to have considerable effect on health outcomes both at personal and population level.[29] posit that a sociological contribution to health research has gain a massive attention among health scholars. Research shows that there are social determinants that contribute to certain health behaviors that might be useful or disastrous. This implies that there needs to be a paradigm shift in the assessment of the factor that influence health behavior and more attention be drawn to the individual responsibility for certain health behaviors [3]. [29] emphasize that social determinants have a significant impact on health outcomes and the discussion should not stop at the biological and psychological processes. [3] acknowledge that education, racism and economic resources are some of the vital social determinants that have not been paid the attention they deserve as they have significant health effects. This study has been able to show that age, gender, income and source of medical information have a significant effect on the health behavior of a person. The study found out that younger people were more likely to take a COVID-19 test compared older ones. Based on gender, females showed a higher likelihood of being tested compared to their male counterparts.

[9] postulates that during a pandemic the mental health of the population plays a vital role in the spread of the disease. The psychological needs during a pandemic are as important as they shape the process of a pandemic management. The adherence of the people to the public health measures such as testing are largely influenced by psychological issues especially, they coping mechanism to threat of infection and loss among the people. In the case of this current pandemic, it is clear that the disease has brought some psychological effects on the people. In this study, it has been found that feelings of emotions such as feeling of being unable to control the course of life during the pandemic, feeling of difficulty to overcome, feeling of constant worry, de-

pression and hopelessness have had a significant effect on people's response to testing for the disease.

As it can be seen in figure 4.3, demographical variables gender, age, income level marital status and race were found to be significant factors the determine a person's choice to take a corona virus test. Education and working status did not show any significant effect. The table shows that those persons who stated that they receive their news about corona virus from their physicians were significantly influenced to take corona virus test. Those who received news from media outlets, social media, health entities, friends and firmly did not show any significant effect on their willingness to take the test. Emotional feelings such as feeling of worrying, nervousness and depression also showed to influence a person choice to take COVID-19 test.

CHAPTER V

CONCLUSION AND FUTURE STUDIES

5.1 Conclusion

Based on the findings of the analysis, various factors have been found to have a significant effect on an individual's choice to take a corona virus test. In the mitigation of the spread of this deadly virus, identifying and quarantining the positive cases is an essential step that has to be widely exercised. Given that individuals are let to take the tests willingly, the results of the test inform them on the right action to take to protect their loved ones. The findings show that both demographic, social and emotional characteristics on an individual have a significant effect on an individual's action to go for a COVID-19 test. The gradient boosting techniques has been able to identify the most significant to the least significant factors. The model shows that age has the highest influence on a person's choice to take the test. From the bivariate analysis, it was identified that there was a significant age difference between individuals who took corona virus test and those who did not. Other demographic factors such as race, gender and income level were also found to have a significant influence. The model also showed that people who obtain news about corona virus from their physicians are more likely to take a corona test.

5.2 Future Studies

This study was conducted using cross-sectional data, which cannot be used in identifying cause-effect relationship. This study has been able to determine the factors that have some correlation with taking COVID-19 test, but these findings cannot be used to imply causation. Future studies should look into conducting a longitudinal study what will be able to show causality between the

significant factors in this study and the intent to take a test of a pandemic. Also, the this study used secondary information. The use of secondary information is a huge limitation in this study as the data was not collected purposefully to answer the objective of this study. In future, similar studies should consider utilizing primary data to ensure that most of the variables in the data are relevant to the study and no proxy variables are used.

BIBLIOGRAPHY

[1] J. B. AGUILAR, J. S. FAUST, L. M. WESTAFER, AND J. B. GUTIERREZ, *Investigating the impact of asymptomatic carriers on covid-19 transmission*, MedRxiv, (2020).

[2] K. J. BOURASSA, D. A. SBARRA, A. CASPI, AND T. E. MOFFITT, *Social distancing as a health behavior: County-level movement in the united states during the covid-19 pandemic is associated with conventional health behaviors*, Annals of Behavioral Medicine, 54 (2020), pp. 548–556.

[3] P. BRAVEMAN, S. EGERTER, AND D. R. WILLIAMS, *The social determinants of health: coming of age*, Annual review of public health, 32 (2011), pp. 381–398.

[4] P. CAMPOS-MERCADE, A. N. MEIER, F. H. SCHNEIDER, AND E. WENGSTRÖM, *Prosociality predicts health behaviors during the covid-19 pandemic*, Journal of public economics, 195 (2021), p. 104367.

[5] W. CAO AND T. LI, *Covid-19: towards understanding of pathogenesis*, Cell research, 30 (2020), pp. 367–369.

[6] F. CARAMELO, N. FERREIRA, AND B. OLIVEIROS, *Estimation of risk factors for covid-19 mortality-preliminary results*, MedRxiv, (2020).

[7] J. E. CAVANAUGH AND A. A. NEATH, *The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements*, Wiley Interdisciplinary Reviews: Computational Statistics, 11 (2019), p. e1460.

[8] L. CONNELLY, *Logistic regression*, Medsurg Nursing, 29 (2020), pp. 353–354.

[9] W. CULLEN, G. GULATI, AND B. KELLY, *Mental health in the covid-19 pandemic*, QJM: An International Journal of Medicine, 113 (2020), pp. 311–312.

[10] A. V. DOROGUSH, V. ERSHOV, AND A. GULIN, *Catboost: gradient boosting with categorical features support*, arXiv preprint arXiv:1810.11363, (2018).

[11] H. EJAZ, A. ALSRHANI, A. ZAFAR, H. JAVED, K. JUNAID, A. E. ABDALLA, K. O. ABOSALIF, Z. AHMED, AND S. YOUNAS, *Covid-19 and comorbidities: Deleterious impact on infected patients*, Journal of infection and public health, (2020).

[12] M. FARBOODI, G. JAROSCH, AND R. SHIMER, *Internal and external effects of social distancing in a pandemic*, Journal of Economic Theory, 196 (2021), p. 105293.

[13] G. HÉKIMIAN, G. LEBRETON, N. BRÉCHOT, C.-E. LUYT, M. SCHMIDT, AND A. COMBES, *Severe pulmonary embolism in covid-19 patients: a call for increased awareness*, Critical Care, 24 (2020), pp. 1–4.

[14] C. HERTZMAN, *Commentary on the symposium: biological embedding, life course development, and the emergence of a new science*, Annual Review of Public Health, 34 (2013), pp. 1–5.

[15] W. HSU, C. CHIANG, AND S. YANG, *The effect of individual factors on health behaviors among college students: the mediating effects of ehealth literacy*, Journal of medical Internet research, 16 (2014), p. e3542.

[16] G. JAMES, D. WITTEN, T. HASTIE, AND R. TIBSHIRANI, *An introduction to statistical learning*, vol. 112, Springer, 2013.

[17] M. JAYAWEERA, H. PERERA, B. GUNAWARDANA, AND J. MANATUNGE, *Transmission of covid-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy*, Environmental research, 188 (2020), p. 109819.

[18] S.-D. LE BON, N. PISARSKI, J. VERBEKE, L. PRUNIER, G. CAVELIER, M.-P. THILL, A. RODRIGUEZ, D. DEQUANTER, J. R. LECHIEN, O. LE BON, ET AL., *Psychophysical evaluation of chemosensory functions 5 weeks after olfactory loss due to covid-19: a prospective cohort study on 72 patients*, European Archives of Oto-Rhino-Laryngology, 278 (2021), pp. 101–108.

[19] X. LI, S. XU, M. YU, K. WANG, Y. TAO, Y. ZHOU, J. SHI, M. ZHOU, B. WU, Z. YANG, ET AL., *Risk factors for severity and mortality in adult covid-19 inpatients in wuhan*, Journal of Allergy and Clinical Immunology, 146 (2020), pp. 110–118.

[20] A. LIAW, M. WIENER, ET AL., *Classification and regression by randomforest*, R news, 2 (2002), pp. 18–22.

[21] Y. LIU, X. CHEN, X. ZOU, AND H. LUO, *A severe-type covid-19 case with prolonged virus shedding*, Journal of the Formosan Medical Association, 119 (2020), p. 1555.

[22] W. MCKIBBIN, R. FERNANDO, ET AL., *The economic impact of covid-19*, Economics in the Time of COVID-19, 45 (2020).

[23] P. M. MUNNOLI, S. NABAPURE, AND G. YESHAVANTH, *Post-covid-19 precautions based on lessons learned from past pandemics: a review*, Journal of Public Health, (2020), pp. 1–9.

[24] U. NIEMANN, B. BOECKING, P. BRUEGGEMANN, W. MEBUS, B. MAZUREK, AND M. SPILIOPOULOU, *Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables*, PloS one, 15 (2020), p. e0228037.

[25] D. OF HEALTH, D. HUMAN SERVICES, WASHINGTON, H. P. . (GROUP), AND U. S. G. P. OFFICE, *Healthy people 2010: Understanding and improving health*, US Department of Health and Human Services, 2000.

[26] A. M. POLLOCK AND J. LANCASTER, *Asymptomatic transmission of covid-19*, 2020.

[27] H. S. RODRIGO, *Bayesian artificial neural networks in health and cybersecurity*, University of South Florida, 2017.

[28] M. E. SELVAN, *Risk factors for death from covid-19*, Nature Reviews Immunology, 20 (2020), pp. 407–407.

[29] S. E. SHORT AND S. MOLLBORN, *Social determinants and health behaviors: conceptual frames and empirical advances*, Current opinion in psychology, 5 (2015), pp. 78–84.

[30] S. SPERANDEI, *Understanding logistic regression analysis*, Biochemia medica, 24 (2014), pp. 12–18.

[31] M. VAN GERWEN, M. ALSEN, C. LITTLE, J. BARLOW, E. GENDEN, L. NAYMAGON, AND D. TREMBLAY, *Risk factors and outcomes of covid-19 in new york city; a retrospective cohort study*, Journal of medical virology, 93 (2021), pp. 907–915.

[32] R. WANG, J. CHEN, Y. HOZUMI, C. YIN, AND G.-W. WEI, *Decoding asymptomatic covid-19 infection and transmission*, The journal of physical chemistry letters, 11 (2020), pp. 10007–10015.

[33] T. WANG, Z. DU, F. ZHU, Z. CAO, Y. AN, Y. GAO, AND B. JIANG, *Comorbidities and multi-organ injuries in the treatment of covid-19*, The Lancet, 395 (2020), p. e52.

[34] P. WATERWORTH, M. PESCUD, R. BRAHAM, J. DIMMOCK, AND M. ROSENBERG, *Factors influencing the health behaviour of indigenous australians: Perspectives from support people*, PloS one, 10 (2015), p. e0142323.

[35] S. H. WOOLF AND P. BRAVEMAN, *Where health disparities begin: the role of social and economic determinants—and why current policies may make matters worse*, Health affairs, 30 (2011), pp. 1852–1859.

[36] Z. ZHENG, F. PENG, B. XU, J. ZHAO, H. LIU, J. PENG, Q. LI, C. JIANG, Y. ZHOU, S. LIU, ET AL., *Risk factors of critical & mortal covid-19 cases: A systematic literature review and meta-analysis*, Journal of infection, 81 (2020), pp. e16–e25.

APPENDIX A

APPENDIX A

1.1 Data Analyses

1.1.1 Variables

- The dependent variable in this study was the coronavirus test status, which consisted of two categories i.e. Tested and Not tested.
- All variables used in the study were categorical except the variable age.
- Chi-square test was used to test association between variables.
- The significance value used in all the analyses performed in this study were done at $\alpha = 0.05$.
- 23 predictor variables were selected for the study.
- 8 demographic variables (Gender, Age, Marital Status, Education, race, working status, immigration status and income)
- 5 Corona related variables (i.e. sources of COVID-19 news).
- 2 health related variables (Having medical insurance and medical condition).
- 8 mental health related variables.

1.1.2 Machine learning models (Classifiers)

- 5 machine learning (ML)models were used.
- These ML classifiers include logistic regression model, CART decision tree model, gradient boosting model, artificial neural network and random forest model.

- These ML classifiers have been found to have strong predictive capabilities hence their use in this study.
- The Random forest model was found to produce the best results in this study.

1.1.3 Data analysis steps

Summary of the data analysis steps

- Analysis was performed through a series of 4 stages.
- The first stage of analysis was the splitting of the sample data into training and testing datset.
- The training dataset was used to train the model using the 5 different classification techniques.
- The second stage of analysis was testing the trained model to evaluate their performance.
- The testing dataset was used to test the model performance.
- The process of training and testing was iterative and the set seed values change with each iteration.
- The best models in each classification technique was selected.
- In the third stage, the best models were used to determine the best predictors of the outcome variable.
- The best of the best model was selected in the fourth step to be used to predict the likelihood of a person to take the COVID-19 test.

Step 1: Classifier training

- The data was first split into training (352/440, 80%) and testing (88/440, 19%) data sets. The classifiers were trained using the training data and their performance evaluated using testing data set.

- For each of the ML classifier, 10 different random initialization were trained and tested.

Step 2: Model Performance Evaluation

- The test data set was used to test the model performance of the fitted classifiers.
- The parameters for the evaluation of model performance that were used in this study include accuracy, sensitivity, specificity and AUC.
- These parameters were represented as averages and standard deviations that were derived from the 10 iterations performed for each technique.
- The best model was the one with the highest AUC value.

Step 3: COVID-19 test predictors

- Using Shapley Additive Explanations (SHAP), a model-agnostic posthoc framework, the predictors of the best model were analyzed.
- SHAP analysis was essential in model interpretation and identification of the most significant factors in the model.
- SHAP analysis compare model predictions that consider or fail to consider specific features in a model, then computes SHAP values that indicate the importance of a feature.

Step 4: Identification of persons that are most likely to take COVID-19 test

- Random Forest was used in the identification of the people that are most likely or least likely to take COVID-19 test.

APPENDIX B

APPENDIX B

Data analysis was performed used R Version 4.1.0

2.0.1 Data Partitioning

```r
library(caret)
set.seed(122)
trainIndex <- createDataPartition(y=samp_covid$Tested_For_Corona,p = 0.8,list =FALSE,times

training<-samp_covid[trainIndex,-6] #Column 6 is the original Race variable. Since we have
testing<-samp_covid[-trainIndex,-6]

dim(training) #352 by 24
dim(testing) # 88 by 24

table(training$Tested_For_Corona)
# Not Tested     Tested
# 264         88

table(testing$Tested_For_Corona)
# Not Tested     Tested
# 66         22
```

Figure B.1: Partitioning of the data

2.0.2 Logistic regression

```r
#****************************************
#***************Logistic Regression*******
#*

#This contains the all the variables 23 independent variables)
glm.model1 <- glm(Tested_For_Corona~.,
                  data =training, family = binomial) %>%stepAIC(trace = FALSE)
summary(glm.model1)
anova(glm.model1,test="Chisq")
anova(glm.model1,test="LRT")

# Make predictions
library(tidyr)
GLM_probabilities <- glm.model1 %>% predict(testing, type = "response")


GLMpredicted.classes <- ifelse(GLM_probabilities > 0.165, "Tested", "Not Tested")
table(testing$Tested_For_Corona,GLMpredicted.classes)
#confusionMatrix(GLMpredicted.classes, testing$Tested_For_Corona)

library(pROC)
roc <- roc(testing$Tested_For_Corona,GLM_probabilities)
roc
```

Figure B.2: Training and testing the logistic regression

2.0.3 Decision tree

```
#*************************************************************************************************
#*************************************1.Decision Tree Models*************************************
#*************************************************************************************************
#To get help
?rpart()
#rparttreeRevised.modelnew is the revised model ran based on all dicotaomus variables (all
#training14 contains the training data with new variables
##Decision Tree on RV1
library(rpart)
set.seed(274)
rparttree.model1<- rpart(Tested_For_Corona~.,training,parms=list(split="gini"),control = rp
rparttree.model1
summary(rparttree.model1)

rparttree.model1$variable.importance

#Plotting the Tree
library(rpart.plot)
rpart.plot(rparttree.model1,type=3,under = TRUE,
           space = 0,tweak=0.8,under.cex = 1,fallen.leaves = TRUE,branch.lty = 3)

rpart.plot(rparttree.model1,type = 3, clip.right.labs = FALSE)

#Making Predictions using Test Data
tree.pred1=predict(rparttree.model1,testing,type="class")
tree.pred1[1:10]

predictions_DT0 <- as.data.frame(predict(rparttree.model1,testing, type = "prob"))
predictions_DT0$predict <- names(predictions_DT0)[1:2][apply(predictions_DT0[,1:2],1, whicl
predictions_DT0$observed <- testing$Tested_For_Corona

roc.DT0 <- roc(ifelse(predictions_DT0$observed=="T", "T", "NT"),
               as.numeric(predictions_DT0$`T`), grid=T, plot=T,
               print.auc =T, col ="red")

roc.DT0 #0.6171
```

Figure B.3: Training and testing the decision tree

2.0.4 Gradient Boosting

```
require(xgboost)
head(training)
nrounds <- 1000
tune_grid <- expand.grid(
  nrounds = seq(from = 200, to = nrounds, by = 50),
  eta = c(0.025, 0.05, 0.1, 0.3),
  max_depth = c(2, 3, 4, 5, 6),
  gamma = 0,
  colsample_bytree = 1,
  min_child_weight = 1,
  subsample = 1
)
dummy_x_training=model.matrix(Tested_For_Corona~.,training)[,-1]
head(dummy_x_training)
class(dummy_x_training)
dummy_y_training<-training$Tested_For_Corona
dummy_x_testing=model.matrix(Tested_For_Corona~.,testing)[,-1]
head(dummy_x_testing)
dummy_y_testing<-testing$Tested_For_Corona
set.seed(4852)
tc = trainControl(method = "repeatedcv", number = 3)
xgb_model1 <- caret::train(
  x = dummy_x_training, y = dummy_y_training,trControl = tc,tuneGrid = tune_grid,
  method = "xgbTree",verbose = TRUE)
summary(xgb_model1)
XGBoost_prediction1 <- predict(xgb_model1, newdata = dummy_x_testing)
XGBoost_prediction1
table(testing$Tested_For_Corona, as.factor(XGBoost_prediction1))
predictions_XGboost1<- as.data.frame(predict(xgb_model1,dummy_x_testing, type = "prob"))
predictions_XGboost1$predict <- names(predictions_XGboost1)[1:2][apply(predictions_XGboost
predictions_XGboost1$observed <- testing$Tested_For_Corona
head(predictions_XGboost1)
library(pROC)
roc.XGBoost1 <- roc(ifelse(predictions_XGboost1$observed=="T", "T", "NT"),
                    as.numeric(predictions_XGboost1$'T'), grid=T, plot=T,
                    print.auc =T, col ="red")
```

Figure B.4: Training and testing of Gradient boosting

2.0.5 Artificial Neural Network

```
#############################################################
##########################5. ANN Analysis ####################
#############################################################



names(training)
dim(training)
dim(testing)

table(training$Tested_For_Corona)
# NT   T
# 264  88

set.seed(7222)
nnfit.model1<- train(training[,-c(8)],training[,8], method = "nnet",
                     trControl = trainControl(method = "cv",number=3), act.fct = "logist

plot(nnfit.model1)


nn.pred2Prob <- predict(nnfit.model1,testing[,-8], type='prob')
head(nn.pred2Prob)
ANN.predCustomClass<-as.factor(ifelse(nn.pred2Prob[,2]>0.151,"T","NT"))
ANN.predCustomClass
conftab <- table(ANN.predCustomClass,testing$Tested_For_Corona)
accuracy(conftab)
specificity(testing$Tested_For_Corona, ANN.predCustomClass)
sensitivity(testing$Tested_For_Corona, ANN.predCustomClass)
nn.pred2.frame<-as.data.frame(nn.pred2Prob)
nn.pred2.prob <- as.numeric(nn.pred2.frame[,2])
nn.roc1<- roc(as.numeric(testing$Tested_For_Corona),nn.pred2.prob, grid=T,
              plot=T, print.auc =T, col ="red")
nn.roc1
```

Figure B.5: Training and testing of the ANN model

2.0.6 Random Forest

```
#### Random Forest

library(caret)
library(randomForest)
library(pROC)


set.seed(12)
rand_f <- randomForest(Tested_For_Corona~.,data=training, proximity = TRUE)
print(rand_f)
pred_rand <- predict(rand_f, testing, type = "prob")
pred_rand <- ifelse(pred_rand[,2]>0.2,"T","NT")
table(pred_rand)
confusionMatrix(pred_rand,testing$Tested_For_Corona)
table(testing$Tested_For_Corona)



rf_roc<- roc(as.numeric(testing$Tested_For_Corona),as.numeric(pred_rand), grid=T,plot=T, p
rf_roc

plot(rand_f)
```

Figure B.6: Training and testing of ANN model

2.0.7 SHAP Analysis for Random Forest

```r
set.seed(42)
library("iml")
library("randomForest")
# data("Boston", package = "MASS")
# rf_boston <- randomForest(medv ~ ., data = Boston, ntree = 50)
#
# XSA <- Boston[which(names(Boston) != "medv")]
# predictor_XSA <- Predictor$new(rf_boston, data = XSA, y = Boston$medv)
#imp <- FeatureImp$new(predictor_XSA, loss = "mae")
library("ggplot2")
plot(imp)
# predictor_XSA2 <- Predictor$new(rf_defaultnew, data = training[,-8], y = training$Tested_
# imp2 <- FeatureImp$new(predictor_XSA2, loss = "ce")
# plot(imp2)
# imp2$results
predictor_XSA3 <- Predictor$new(boost.model_best, data = training[,-8], y = training$Teste
imp3 <- FeatureImp$new(predictor_XSA3, loss = "ce")
imp3
plot(imp3)
imp3$results
#* Check the error
# predictor_XSA3XGBoost <- Predictor$new(xgb_model1, data = as.data.frame(dummy_x_training.
# imp3_XGBoost <- FeatureImp$new(predictor_XSA3XGBoost, loss = "ce")
# plot(imp3_XGBoost )
# imp3_XGBoost $results
predictor_XSA4 <- Predictor$new(rparttree.model1, data = training[,-8], y = training$Teste
imp4 <- FeatureImp$new(predictor_XSA4, loss = "ce")
imp4
plot(imp4)
imp4$results
predictor_XSA5 <- Predictor$new(nnfit.model1, data = training[,-8], y = training$Tested_Fo
imp5 <- FeatureImp$new(predictor_XSA5, loss = "ce")
imp5
plot(imp5)
imp5$results
#*****************************************************************************
```

Figure B.7: SHAP analysis for the Random Forest Model

BIOGRAPHICAL SKETCH

Sheila Rutto is born and raised in Rift-Valley Kenya. She graduated from the University of Texas Rio Grande Valley with a Master of Science in Applied Statistics and Data Science in December 2021. A Graduate Teaching Assistantship recipient, she taught undergraduate mathematics courses during her graduate studies and participated in the Join Statistics Meeting in August 2021, which focused on using machine learning techniques on Covid -19 data. Her future goals are to relate with Actuarial Science with her statistical knowledge and to pursue a Doctorate degree in Interdisciplinary Studies in Mathematics and Statistics. Sheila is reachable at sheilarutto@gmail.com