Classification vs Regression Models for Creating prediction models for vaccination rates of incoming kindergarten students

Chen Yuan Lee
clee94@gmu.edu

Erick Orellana
eorellan@gmu.edu

Kening Zhu
kzhu5@gmu.edu

***Abstract*—Vaccinations have been proven over the years to help treat some of the deadliest diseases as well as to prevent the further spread of those diseases. A major factor that can dictate the success or failure of a vaccine at treating the broader population is to reach heard immunity. To reach heard immunity, it is important to have a certain percentage of the population vaccinated so that they themselves will not be impacted by the disease and so that they will not spread throughout the community. Public health officials in conjunction with education officials understand the importance of immunizations and therefore set standards on vaccination rates and collect data to drive efforts to improve vaccination rates. The state of California is one of those states that has set standards for the vaccinations that kindergarten students should have before entering the school system and collect this information and make it public each year. There is an interest in using this information to make models that can be used to make prediction about vaccination levels at the county level. The most prominent approach taken is a geospatial approach of using a physical map to show vaccination rates. This is a useful visual, but it does not too much in the way of explaining why those vaccination rates are what they are based on certain factors. These geospatial models also do not provide a way to predict how changes in certain factors will impact the vaccination rates. In this study, variables that may impact vaccination rates are explored to generate a regression model and two classification models to understand if those models can be accurately used with the given predictor variables to gage potential changes to vaccination rates based on those given variables.**

***Index Terms*—classification, decision trees, regression models, Machine learning**

I. Introduction

A. Background Information

The main purpose of vaccinations is to help the immune system form protection against diseases [2]. It helps prevent infectious diseases when stimulating the body's adaptive immunity. Babies are born with an immune system that can fight most bacteria, but there are some serious diseases that babies cannot combat naturally. Babies need vaccines to strengthen their immune system. As babies grow and mature into infants and school-aged children, there are vaccines that are recommended to be administered to enhance their immunity to potentially dangerous infectious diseases. There are vaccines that at times may come with adverse effects, however, most doctors agree that the benefits of vaccinations far outweigh any risks that come from them. While vaccines have the potential to save individuals from life threatening diseases the other added benefit is that it can lead to herd immunity. Herd immunity means that the community, or population, can become immune to diseases which can stop the spread of these deadly diseases before they become serious community health issue [2].

The World Health Organization has estimated that in the year 2020, a total of 23 million children world-wide missed some vaccinations, while an estimated 17 million children were not vaccinated at all [5]. The same study uncovered that vaccination rates that combat diseases such as diphtheria, tetanus, pertussis, measles, and polio have stagnated at 86% of children receiving them world-wide [5]. This is alarming because this data shows that many children and communities are at risk of catching and transmitting diseases for which vaccinations already exist that can prevent them. In many countries such as the United States, school systems require school-aged children to be vaccinated before entering the school system as kindergarteners. This is done in effort to drive up the number of vaccinated children in the community as to prevent the issues that come from having an unvaccinated population.

One such example of a state in the United States that requires all children attending public and private schools to receive certain immunizations before entering the school system is California. According to the information provided by the California Department of Health, the vaccines that are required for kindergarten students to have to be considered fully vaccinated are vaccines against Polio, Diphtheria, Tetanus and Pertussis (DTaP), Measles, Mumps and Rubella (MMR), Hepatitis B, and Varicella (Chickenpox) [1]. Although these vaccines are required, not all children entering the school system have them. In these instances, cases are reviewed to understand if the reason for this is something that would warrant an exception such as pre-existing health conditions, the child started the sequence of vaccinations late, or the parents of the children citing personal reasons for not vaccinating their children.

B. Points of Interest

The state of California requires schools to submit vaccination information for all students entering kindergarten at the start of this school year. This vaccination information is analyzed and interpreted to get an understanding of how

efforts to increase vaccination in children across the state are going. Currently, a lot of studies and models that are created focus on geospatial data and generating things such as maps with clusters showing vaccination levels. This creates an issue of not knowing what other variables within those geographic areas may be impacting vaccination efforts. Our study intends to assess the following questions:

1) How effective can a multilinear regression model be at predicting the percentage of kindergarten students fully vaccinated per county in California at the start of the school year? : The first approach taken in this study to formulate a way to predict vaccination rates based on certain key factors is to create a linear regression model. In this case, since there are multiple predictor variables potentially acting on the target variable(vaccination rates), a multilinear regression approach has been selected for further analysis.

2) How effective can a decision tree or a k-nearest neighbor model be at categorizing whether a California county's kindergarten population will be at least 95% vaccinated at the start of the school year?: The second approach taken in this study is to look at the target variable (vaccination rates) as a classification rather than a continuous number. This approach would still provide information on how different predictor variables play a role in being able to predict future vaccination rates. One of the classification methods chosen for further study is decision trees with the intent of providing a more visual method to predict if a county will be adequately vaccinated or not. The other classification method selected is k-nearest neighbors to see how the change of different variables may impact vaccination rates.

3) Is a multilinear regression model, a decision tree model, or a k-nearest neighbor model the most effective way to make predictions about the percentage of all kindergarteners that will be fully vaccinated in the counties of California?: Lastly, if a multilinear regression model, a decision tree, and a k-nearest neighbor model can all be generated for this data, it is important to understand how the three methods compare to each other in terms of accuracy. After understanding how the methods compare to each other, this study will assess how the most accurate method compares to the current geospatial models that exist.

C. Impact of this Study

This project is important because it has the potential of being used for public health and other community outreach initiatives. For example, if it is predicted that the size of the population of a county in California will change, how might that impact the vaccination status of kindergarten students? Or maybe, if the average annual household income is expected to increase by a certain percentage in a county can a portion of the vaccination resources from that county be moved to another? Having predictive models in place could help answer these questions to ensure that resources are being used in the most effective way. Also, if it is determined that the models generated in this paper do not produce the most useful predictions then it can be used to explain why those models do not work so that others working on creating a similar model know to use other methods or can build on what is proposed in this study.

II. Literature review

The subject of vaccinations is a topic of discussion that has come to the forefront of discourse due to the COVID-19 pandemic and the various vaccines that have been developed to reduce the spread and the negative health effects of COVID-19. There are often times discussions or even debates regarding whether to get vaccinated or not, but the conversation goes past even the COVID-19 vaccine to a more general discussion about vaccines, especially the ones that infants and children get before they are old enough to go to school. According to Paul Delamater, "The refusal or delay of childhood vaccinations has been identified in both the popular press and scientific literature as an increasing public health concern across the United States" [3]. The refusal to vaccinate children has negative effects that impact the child and the community at large. As Andrea Praticò explains, vaccinating children is not only about protecting their own health and improving their quality of life but it is also about protecting the community, especially high-risk individuals, from getting deadly diseases for which vaccines have been developed [10]. For this reason, may states across the United States mandate that certain vaccines should be given to children before starting kindergarten although they leave room for exceptions.Machine learning analysis of online text data shows this facts [1]–[6].

The state of California is one that requires that students be vaccinated before starting kindergarten while also leaving room for exceptions. There has been research done broadly into reasons why exceptions are granted and why some parents refuse to have their children vaccinated. For example, the article titled "A systematic review of factors affecting vaccine uptake in young children" explains that there are religious reasons or the potential for allergic or otherwise negative reactions to a vaccine by a child which warrants the need to provide exceptions to being vaccinated [12]. As well as Emily R. Zier mentioned that "Parents are concerned about vaccine ingredients, the number of vaccines on the recommended schedule, the (scientifically discredited) notion that vaccines are linked to autism, and the debate over the necessity of vaccines." [4]. There has also been research done which revealed more personal reasons as to why some parents choose not to vaccinate their children such as misinformation or mistrust on the impacts of the vaccine or simply being in geographic regions that make them unable to reach the available resources to get vaccinated [9]. Being able to allocate resources or distribute information in the most appropriate manner could help increase the number of children that are vaccinated which would help communities be healthier. Similarly, a free vaccination policy was found strongly correlated with higher vaccination intention. Our study findings urge the need to

offer free influenza vaccines to children and more education to parents in order to increase the vaccine uptake rate. [14].

One way in which resource allocation or information distribution regarding vaccination in children could be improved is by creating models to predict or access how certain factors impact the percentage of students that are fully vaccinated. For instance, Robert M. Kaplana pointed that "The survey was cross sectional by design and used a multistage cluster sampling procedure. A total of 1,927 mothers with children of 12–23 months of age were extracted from the children's dataset. Mothers' self-reported data and observations of vaccination cards were used to determine vaccine coverage. An adjusted odds ratio (AOR) with 95% confidence intervals (CI) was used to outline the independent predictors." [11]. There are studies conducted which focus on using geospatial data to understand where the most unvaccinated children are and to make predictions of how shifts in population impact those numbers. The author Tracy Lieu uses geospatial data to create geographic clusters on a map of California to show where different levels of vaccinated students reside [6]. Moreover, the author Paul Delamater builds on Tracy Lieu's concept by building maps that show how mobility of people shifts the level of vaccinated students based on different geographic regions. In addition to, Nita Bharti combines satellite-derived measurements of fluctuations in population distribution with high-resolution measles case reports to develop a dynamic machine learning model [**?**], [**?**], [6]–[36] that illustrates the potential improvement in vaccination campaign coverage if planners account for predictable population fluctuations. [8]. Focusing on geographic regions can aid in allocating resources, but there are factors other than geography which may impact the levels of vaccination.

The author Louise-Anne McNutt published an article titled "Affluence as a predictor of vaccine Reflection and under immunization in California private kindergartens" in which factors that may help to explain the distribution of vaccination rates in geospatial reports are discussed in more detail. This study looks at factors such as cost of tuition at private schools, religious affiliation, and enrollment numbers to analyze vaccination numbers [7]. This article creates opportunities to look beyond geospatial attributes and to consider other socioeconomic factors that may impact vaccination numbers. This also opens opportunities to look into models that do not necessarily generate maps or other geospatial models to study vaccination rates.

III. Materials and Methods

A. Dataset

For this study we will first compile raw data by combining information from the California Government's open data portal and the United States Census Bureau's website. After this we will complete data clean up and begin exploratory data analysis. This will be done to create visualizations and some preliminary calculations to begin to isolate variables and see points of interest.

The dataset titled "2019-2020 Kindergarten Immunizations" forms the foundation of this study. This information is made available on the California open data portal. The source of the data is the public and private schools throughout the state which are required by law to report the immunization status of the kindergarten students at the start of the school year. From this dataset the following fields, or variables, will be used:

- "COUNTY": California county in which the school is located
- "CATEGORY": Label for the vaccination status being reported
- "PERCENT": Percent of students fully vaccinated (Target Variable) Component heads identify the different components of your paper and are not [37]

The second data set from which information will be gathered for the exploratory data analysis stage is an interactive portal from the United States Census Burau titled "QuickFacts." The United States Census Bureau compiles this information and makes it available for public use. We will be using data for the different counties in California from 2019 since the kindergarten data being gathered is from 2019. From this dataset the following fields will be used:

- "Population Estimates": Number of residents in the county
- "Female persons, percent": Percent of population that is female
- "White alone, percent": Used to calculate the percentage of the population belonging to a minority group (100 – "White alone, percentage")
- "High school graduate, or higher": Percent of the population older than 25 years old with a high school diploma or equivalent
- "Persons without health insurance, under age 65": Percentage of the population under 65 years of age with no health insurance
- "Median household income": Median income for the household in 2019 dollars
- "Persons in poverty, percent": Percentage of the population living in poverty
- "Unemployment rate": Percentage of the population unemployed

B. Methods

After completing the preliminary analysis, we will move into creating a multilinear regression model, a decision tree, and a k-nearest neighbor modelto see if a model can be built to predict based on different variables if the average percentage of student's fully vaccinated is at least 95%. We propose to use a portion of the data to train the different models then a portion of the data to test. For the multi-linear regression model the target variable will be a continuous number depicting the percent of kindergarten students that are fully vaccinated. For the classification models the target variable of percentage of vaccinated students will be split into two categories: "95% vaccinated or higher" and the other category will be "Less

than 95% vaccinated." This change to the dataset will be made to accommodate working with linear regression models versus working with classification models. Lastly, we will compare the accuracy of these models and see which is the most accurate and ultimately if the models are useful.

The data collection and initial clean-up will take place in Microsoft Excel. The multilinear regression model will be built and analyzed using the programming language R and the program R-Studio. There will also be a step-wise reduction conducted on the multi-linear regression model generated to see if there are variables that are not contributing in a statistically significant way to the model to see if it can be reasonably improved. The decision tree and the K-Nearest Neighbor model will be built using the program language Python and the program Jupyter Notebook. Python packages such as numpy, pandas, and sklearn will be used to work with the dataset.

To evaluate the different models that we create we intend to first split our dataset into a training and a test data set by randomly splitting up the records 75% for training the models and 25% for testing the models. This will be useful to get an understanding of how our model does at predicting the outcomes of the test data set. The following statistical tests will be used to evaluate the multi-linear regression model: root mean squared error, r-squared, and p-value. To evaluate the accuracy of the decision tree and k-nearest neighbor models a 10-fold cross validation will be used. These statistical values will then be used to formulate a discussion about the accuracy of the models.

IV. Results

The attributes that were chosen for these models are the total population, percent of the population that is female, percent of the population that belongs to a minority race or ethnic group, the percent of the population that graduated high school, percent of the population under 65 years of age that does not have health insurance, median income for the county, percent of the population that is unemployed, and percent of the population in poverty. These eight attributes were chosen based on literature research as factors that may contribute to parents getting their school age children vaccinated and therefore monitoring the changes of these variables could be useful in making predictions around the overall rate of vaccination.

A. Classification: Decision Tree

The dataset was formatted so that the target variable "PER_FULL_VAX" (representing the percent of kindergarten children fully vaccinated) displayed either a 1 if the value is at or above a 95% and a 0 if the value is below a 95%.

The decision tree that is depicted in figure three was created using the Sklearn's "DecisionTreeClassifier" function. The decision tree begins with looking at the percent of population that is living in poverty and moves down from there based on certain criteria on different attributes. There is no additional pruning done to the tree.This decision tree was created using 75% of the data to train the model. When the model is tested using the remaining 25% of the data, it is revealed that the accuracy in classifying whether a county would have at least 95% of the kindergarten population vaccinated was approximately 66.7Ṫhis value is fairly low which does not present much confidence that this model would be useful in making decisions about things such as resource distribution or making public health policies regarding vaccinations.

In an effort to dive a bit deeper into the decision tree model, Sklearn's "DecisionTreeClassifier" was ran without using test and training data. A model was created using all of the data and a 10-fold cross validation to measure accuracy. This resulted in an accuracy of 61.9% which was a decrease in accuracy when compared to the decision tree made with a 75% train / 25% test split in the data. Therefore, the first decision tree model performed better but it still not useful enough in the context of predicting how different factors may impact vaccination rates of counties in California.

To extend on decision trees, we took a look at random forests still using Sklearn's "RandomForestClassifier" in Jupyter Notebook. An estimator value of 10 was used for this exercise and the model generated had an accuracy of approximately 62.00% which is also a lower accuracy than the decision tree that was generated using a test/train split dataset.

B. Classification: K-Nearest Neighbor

Once again, the dataset was formatted so that the target variable "PER_FULL_VAX" (representing the percent of kindergarten children fully vaccinated) displayed either a 1 if the value is at or above a 95% and a 0 if the value is below a 95%.

Another classification model that could be used to predict a county in California will have at least 95% of school age children vaccinated is K-Nearest Neighbors. Sklearn's "KNeighborsClassifier" was used to create models for this classification. There were tests made using different values for the 11n_neighbors" variable ranging from 1 to 10 to see which produced the best accuracy. The idea behind testing a range from 1 to 10 is to see if increasing the variable would also increase the accuracy. the accuracy of these models was then calculated using a 10-fold cross validation .

This test proved that for this model, increasing the n_neighbors value did not necessarily cause an increase or a decrease in accuracy. The accuracies kept increasing and decreasing without a discernable trend that could be seen at a quick glance. When looking at the summary of accuracies in figure five, it becomes evident that the highest level of accuracy was achieved at an "n_neighbors" value of 5. The accuracy at that point was about 72%. This is measurement of accuracy comes as an improvement over the decision tree and random forest models but it still does not spark enough confidence in using this model as a replacement over the traditional geospatial models.

C. Multi-Linear Regression

To begin the process of building a multi-linear regression model the p-value of all predictor variables in relation to the predictor variable were calculated. A hypothesis test was then conducted using these p-values shown in figure six with a significant value of 0.05. The following hypothesis was tested:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned} \tag{1}$$

Based on this test it is apparent that the variables "PERCENT_MINORITY" and "PERCENT_HS_GRAD" are the only variables with a p-value less than 0.05 leading to the conclusion that there is strong evidence that these predictor variables on their own are statistically useful in predicting the vaccination rates. The two variables were graphed to visually inspect the results of the hypothesis test. These two graphs can be seen in figure seven and in figure eight:

However, the majority of the predictor variables chosen for this study do not seem to be statistically significant in creating a linear regression model for this prediction.

After validating that there exists evidence that points to some of the predictor variables being of statistical significance, a multi-linear regression model was created. The following summary statistics were also generated with this model:

Key values evaluated to see the usefulness of this model for predicting the vaccination rate for kindergarten students in California include Multiple R-Squared, Adjusted R-Squared, and p-value. Looking at the summary statistics in figure nine, it is shown that the Multiple R-Squared the Adjusted R-Squared are both very low. This would imply that the predictive multi-linear regression model created using all predictor variables is not statistically useful in predicting or explaining the variation in the dependent variable. This is further confirmed when looking at the overall p-value for the model. The p-value is well above the threshold set at 0.05 also showing that this model is not of statistical significance in predicting the target variable.

In an effort to do further analysis on a multi-linear regression model, the model with every single predictor variable in it was taken and a stepwise regression was completed on it. This decision came about to uncover if potentially removing any individual, or combination, of predictor variables that are of less statistical significance would produce a more accurate multi-linear regression model. Similar to the multi-linear regression model that has been built, the backwards step-wise regression was computed using the language R in R-Studio. The original dataset was also split into a training set and a test set to be able to gather information about the accuracy of the model that would be generated using the backwards step-wise regression. The model was then optimized and summary statistics were computer for the model that had been generated. figure ten displays the summary of results for computing the stepwise regression:

The results of running a backwards stepwise regression eliminated seven of the eight predictor variables from the original multi-linear regression model. According to the analysis of running a backward step-wise regression technique on the training dataset, the model to predict the vaccination rate of kindergarten students in California would be the following:

$$Y = \beta_0 + \beta_1 X_1 \tag{2}$$

- Y is the predicted percent of kindergarten students that are vaccinated in California
- X_1 is the percent of the population for a county that has graduated high school

While this "optimizes" the prediction model, based on literature search it is highly unlikely that a model with only one predictor variable can help to predict vaccination rates for a county in California. This appears to be a case of overfitting the data. Therefore, no further summary statistics were running on the model that was generated using the backwards stepwise regression. For now, it appears as though the regression model with all predictor variables is the most statistically useful multi-linear regression model that can be generated with the provided data.

D. Summary of Findings

Decision Tree:

- Low accuracy, not an effective predictive model for the rate of vaccinated kindergarten students per county in California.

K-Nearest Neighbor:

- Low accuracy, not an effective predictive model for the rate of vaccinated kindergarten students per county in California.

Multi-linear regression:

- Low accuracy and not enough statistical evidence to prove that this model is effective in predicting the rate of vaccinated kindergarten students per county in California.

V. Limitaioans and next steps

A. Limitations

One limitation of this study is the frequency in which the United States' Census Bureau collects and publishes information related to factors such as population sizes and levels of education or poverty. Also, the United States' Census Bureau relies on self-reporting which may not always be fully accurate if there is a misunderstanding on the form that is being filled out or if the people filling it out feel uncomfortable with providing accurate information for certain questions. The state of California publishes vaccination information every year but information for the United States' Census is not published annually which causes gaps in gathering data for years outside of the United States Census. The state of California tracks some of the same information as the United States Census on an annual basis, but they do not make as much of it public as the United States Census bureau does. Using both the California Census data and the United States Census Bureau's data to gather the same data may lead to inconsistencies or discrepancies if the information is not collected in the same

way for example. This becomes a limitation on getting all of the data from the same sources every year.

A second limitation to this report is that there is not a published target for vaccination rate. For this study, a rate of 95% was selected as the target vaccination rate based on literature review on herd immunity levels for the different vaccines that are required for kindergarten students in the state of California. For the majority of the vaccines that are required by kindergarten students, the reported target for herd immunity is 95% and therefore that is what was selected as the target in this report. If the target is to be defined to be something different, this study is setup in a way where the new target variable could be set, and the models could be generated once again to test the different regression and classification models.

Finally, a third limitation is that this report focuses on data from the year 2019 but there have been major shifts in public health since then due to the global COVID-19 pandemic. Since the rise of the global pandemic there has been a higher discourse around vaccinations and shifting attitudes by the public around vaccines. There is also public discourse surrounding states such as California potentially requiring the COVID-19 vaccine as one of the vaccines that are needed for a student to be considered "Fully Vaccinated." This is to say that further study may be needed to see if these models and predictors based on pre-COVID-19 data hold true past the changes that the COVID-19 virus brought to the public health space.

B. Next Steps

This project still has the potential to create regression or classification models that can be used in generating predictions about vaccination rates in California. One step would be to identify other variables that may be statistically significant in predicting vaccination rates in the state. This could include factors such as religious affiliation or others that impact a parent's decision to get a child vaccinated before entering the school system in kindergarten. It could be that other factors could be more impactful predictor variables for vaccination rates than the ones that were studied in this report.

Another route that this report could take is breaking down the data further geographically. In this report the data which is being studied is at the county level. It could be that diving in deeper and looking at data to something like the city level could provide better information. Some counties are larger than others and may have larger levels of disparity in sectors such as graduation rates or average household income. Looking at a smaller geographic region such as city may result in lower levels of disparity. Also, this would generate more data points which may be better for training the predictive models as long as it does not lead to overfitting the models.

As hinted to in the limitations section, another potential next step would be to study in how the global COVID-19 pandemic has impacted these models. The COVID-19 pandemic has had economic, social, and even political impacts since the start of the 2019 school year which is when the data was taken for this report. It is difficult to study current and future public health trends and models without brining into consideration the potential impacts of the pandemic. The pandemic may have created new factors to consider when predicting vaccination rates that may be better statistically correlated than what was known or studied prior to the pandemic.

When it comes to decision tree model, it may be beneficial to apply pruning to the tree to uncover if that creates better accuracy in the model without leading to overfitting the model. For example, with the multi-linear regression model that was generated, there was also a backwards stepwise regression completed on the model to understand if removing certain predictor variables could improve accuracy. With the stepwis-eregression it did lead to overfitting with relying on only one variable to make a prediction, but even this provided useful information to know that even excluding the less statistically significant variables still did not improve the accuracy of the model that was generated.

VI. Discussion and Conclusion

A. Discussion Around Guiding Questions

1) How effective can a multilinear regression model be at predicting the percentage of kindergarten students fully vaccinated per county in California at the start of the school year?

   Based on our studies in this report, it was uncovered that a multi-linear regression model based on the data and factors considered does not provide an accurate way to predict vaccination rates based on p-value and error rate evaluation. Even when a stepwise regression was completed to remove the less statistically significant variables, the model that was created was not useful when considering the complexity of the issue being considered.
2) How effective can a decision tree or a k-nearest neighbor model be at categorizing whether a California county's kindergarten population will be at least 95% vaccinated at the start of the school year?

   This study uncovered that based on the data was used and the variables that were studied, the classification methods of k-nearest neighbor and decision trees did not provide accurate models for predicting vaccination rates in California. Even when further analysis was conducted by looking into random forests or a range of values for the k-nearest neighbor model, the accuracy of the models was not strong enough to suggest that the models could be statistically useful in predicting vaccination rates in California.
3) Is a multilinear regression model, a decision tree model, or a k-nearest neighbor model the most effective way to make predictions about the percentage of all kindergarteners that will be fully vaccinated in the counties of California?

When comparing all of the models that were studied for this report in regard to Classification and Regression, the model that performed the best is K-nearest neighbors. However, even though this model outperformed the other's studied, it still did not provide an acceptable level of accuracy to propose it as an effective way to predict vaccination rates in the state of California.

B. Conclusion

The current standard for tracking and displaying information related to vaccination rates in the state of California is creating and presenting geospatial maps that display a distribution of vaccination rates across different geographical areas. These maps vary, with some appearing like heat-maps across the geographic location and others using bubbles or a different graphic on a map to provide information related to vaccination rates. These maps are useful in understanding what areas may have high or low vaccination rates but they do not necessarily give an explanation for those rates. They also do not provide much information in the way of predicting what changing certain social, economic, or political factors in the area will do. This information is useful for public health officials to better prepare for changes that may negatively impact vaccination rates. In an effort to fill that gap and provide a form to make predictions on vaccination rates based on predictor variables, this report took on looking at classification methods such as decision trees, random forests, and k-nearest neighbor as well as a multi-linear regression model.

Based on the data that was used to create the machine learning methods described in this study, at this time none of them provide great confidence that they can provide accurate predictions that could be useful to make public health decisions around vaccinations. With the classification methods, the highest accuracy achieved was approximately 72% which does not hit the proposed target of approximately 90% accuracy. With the multi-linear regression model that was generated, p-value and error rates also did not instill statistical confidence that the models generated could be useful in making predictions about how changes in the predictor variables impact vaccination rates.

At this time, it is our conclusion that the current geospatial methods should continue to be used as the primary form of providing information related to vaccination rates in the state of California. In order for these machine learning methods that were explored in this report to be useful further research would need to be conducted to find more accurate predictor variables or datasets that provide more insight into the behaviors and circumstances that impact vaccination rates. There are also other machine learning methods for making predictions or for classification that could be explored such as naïve-bayes and others that may use similar strategies to the ones in this report and may have better results.

REFERENCES

[1] J. H. Fetzer, "Disinformation: The use of false information," vol. 14, no. 2, pp. 231–240, May 2004. [Online]. Available: https://doi.org/10.1023/b:mind.0000021683.28604.5b

[2] M. Fernandez and H. Alani, "Online misinformation." ACM Press, 2018. [Online]. Available: https://doi.org/10.1145/3184558.3188730

[3] M. Heidari, J. H. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2020, pp. 480–487.

[4] H. Zhang, A. Kuhnle, J. D. Smith, and M. T. Thai, "Fight under uncertainty: Restraining misinformation and pushing out the truth." IEEE, Aug. 2018. [Online]. Available: https://doi.org/10.1109/asonam.2018.8508402

[5] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2020, pp. 1–6.

[6] W.-Y. S. Chou, A. Oh, and W. M. P. Klein, "Addressing health-related misinformation on social media," vol. 320, no. 23, p. 2417, Dec. 2018. [Online]. Available: https://doi.org/10.1001/jama.2018.16865

[7] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020, pp. 322–329.

[8] L. Cui and D. Lee, "Coaid: COVID-19 healthcare misinformation dataset," *CoRR*, vol. abs/2006.00885, 2020. [Online]. Available: https://arxiv.org/abs/2006.00885

[9] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–6.

[10] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," vol. 359, no. 6380, pp. 1094–1096, Mar. 2018. [Online]. Available: https://doi.org/10.1126/science.aao2998

[11] Q. Su, M. Wan, X. Liu, and C.-R. Huang, "Motivations, methods and metrics of misinformation detection: An NLP perspective," vol. 1, no. 1-2, p. 1, 2020. [Online]. Available: https://doi.org/10.2991/nlpr.d.200522.001

[12] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, 2017, pp. 963–972. [Online]. Available: https://doi.org/10.1145/3041021.3055135

[13] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a specific business domain," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.

[14] C. Yang, R. C. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013. [Online]. Available: https://doi.org/10.1109/TIFS.2013.2267732

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[16] D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport, "Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection," *Nature Medicine*, vol. 27, no. 7, pp. 1205–1211, May 2021. [Online]. Available: https://doi.org/10.1038/s41591-021-01377-8

[17] J. Havey. "pharma research progress hope,". [Online]. Available: https://catalyst.phrma.org/a-year-and-a-half-later-the-biopharmaceutical-industry-remains-committed-to-beating-covid-19?utm_campaign=2021-q3-cov-inn&utm_medium=pai_srh_cpc-ggl-adf&utm_source=ggl&utm_content=clk-pol-tpv_scl-geo_std-usa-dca-pai_srh_cpc-ggl-

[18] M. Heidari, H. James Jr, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.

[19] P. R. Krause, T. R. Fleming, R. Peto, I. M. Longini, J. P. Figueroa, J. A. C. Sterne, A. Cravioto, H. Rees, J. P. T. Higgins, I. Boutron, H. Pan, M. F. Gruber, N. Arora, F. Kazi, R. Gaspar, S. Swaminathan, M. J. Ryan, and A.-M. Henao-Restrepo, "Considerations in boosting COVID-19 vaccine immune responses," *The Lancet*, vol. 398, no. 10308, pp. 1377–1380, Oct. 2021. [Online]. Available: https://doi.org/10.1016/s0140-6736(21)02046-8

[20] J. H. Kim, F. Marks, and J. D. Clemens, "Looking beyond COVID-19 vaccine phase 3 trials," *Nature Medicine*, vol. 27, no. 2, pp. 205–211, Jan. 2021. [Online]. Available: https://doi.org/10.1038/s41591-021-01230-y

[21] E. C. Fernández and L. Y. Zhu, "Racing to immunity: Journey to a COVID-19 vaccine and lessons for the future," *British Journal of Clinical Pharmacology*, vol. 87, no. 9, pp. 3408–3424, Jan. 2021. [Online]. Available: https://doi.org/10.1111/bcp.14686

[22] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2020, pp. 0542–0547.

[23] J. Howard. "could covid-19 vaccine boosters be necessary? here's what experts are saying". [Accessed October 2021]. [Online]. Available: https://www.wesh.com/article/could-covid-19-vaccine-boosters-be-necessary-heres-what-experts-are-saying/36519793

[24] A. Ain, "The WHO is right to call a temporary halt to COVID vaccine boosters," *Nature*, vol. 596, no. 7872, pp. 317–317, Aug. 2021. [Online]. Available: https://doi.org/10.1038/d41586-021-02219-w

[25] E. Callaway, "COVID vaccine boosters: the most important questions," *Nature*, vol. 596, no. 7871, pp. 178–180, Aug. 2021. [Online]. Available: https://doi.org/10.1038/d41586-021-02158-6

[26] A. Weatherton. "health expert says booster shot could be needed after getting covid-19 vaccine". [Accessed June 8, 2021]. [Online]. Available: https://www.13newsnow.com/article/life/booster-shot-may-be-needed-after-covid-19-vaccine/291-49a8966c-3d91-48ad-99a0-02905c5593cc

[27] M. Heidari, S. Zad, M. Malekzadeh, P. Hajibabaee, , S. HekmatiAthar, O. Uzuner, and J. H. J. Jones, "Bert model for fake news detection based on social bot activities in the covid-19 pandemic," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.

[28] P. Naaber, L. Tserel, K. Kangro, E. Sepp, V. Jürjenson, A. Adamson, L. Haljasmägi, A. P. Rumm, R. Maruste, J. Kärner, J. M. Gerhold, A. Planken, M. Ustav, K. Kisand, and P. Peterson, "Dynamics of antibody response to BNT162b2 vaccine after six months: a longitudinal prospective study," *The Lancet Regional Health - Europe*, vol. 10, p. 100208, Nov. 2021. [Online]. Available: https://doi.org/10.1016/j.lanepe.2021.100208

[29] S. J. Thomas, E. D. Moreira, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. P. Marc, F. P. Polack, C. Zerbini, R. Bailey, K. A. Swanson, X. Xu, S. Roychoudhury, K. Koury, S. Bouguermouh, W. V. Kalina, D. Cooper, R. W. Frenck, L. L. Hammitt, Özlem Türeci, H. Nell, A. Schaefer, S. Ünal, Q. Yang, P. Liberator, D. B. Tresnan, S. Mather, P. R. Dormitzer, U. Şahin, W. C. Gruber, and K. U. Jansen, "Safety and efficacy of the BNT162b2 mRNA covid-19 vaccine through 6 months," *New England Journal of Medicine*, vol. 385, no. 19, pp. 1761–1773, Nov. 2021. [Online]. Available: https://doi.org/10.1056/nejmoa2110345

[30] E. Dolgin, "COVID vaccine immunity is waning — how much does that matter?" *Nature*, vol. 597, no. 7878, pp. 606–607, Sep. 2021. [Online]. Available: https://doi.org/10.1038/d41586-021-02532-4

[31] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0285–0291.

[32] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0255–0261.

[33] "virginia open data portal". [Accessed November 6, 2021]. [Online]. Available: https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-Vaccines-DosesAdmini/28k2-x2rj

[34] U.S. Food and Drug Administration. "covid-19 vaccines". [Accessed November 6, 2021]. [Online]. Available: https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines

[35] populationU. "populationU". [Accessed October 10, 2021]. [Online]. Available: https://www.populationu.com/us/virginia-population

[36] M. Malekzadeh, P. Hajibabaee, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, "Review of graph neural network in text classification," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.

[37] S. Akon and A. Bhuiyan, "Covid-19: Rumors and youth vulnerabilities in bangladesh," 07 2020.