Covid-19 Vaccine Hesitancy Prediction using Machine Learning Techniques

Leyon Ibn Kamal
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
leyon.kamal@northsouth.edu

Mubassir Jahan
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mubassir.jahan@northsouth.edu

Md. Faisal
Departmennt of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
md.faisal3@northsouth.edu

Dr. Sifat Momen
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
sifat.momen@northsouth.edu

***Abstract*—COVID-19 pandemic caused lots of losses, both human and economic. To bring the amount of COVID cases down vaccination of populations is crutial. One of the problems when it comes to vaccinating people, is that many are hesitant to get vaccinated for various reasons. So if we could predict those who might be hesitant to get vaccinated, then there could be an effort made to help reduce hesitancy. Our paper aims to do just that. We created a predictive model that takes advantage of data from previous work where they collected data from volunteers regarding how they percieve their health care system and vaccination in general. By using a machine learning model to predict whether a user is hesitant to get vaccinated or not, we can identify the ones that may be hesitant automatically. T**

***Index Terms*—vaccine, COVID-19, hesitancy,machine learning, xgboost, SVM, logistic Regression, Decision tree, Random Forest**

I. Introduction

More than a year since its declaration of Coronavirus Disease 2019 (COVID-19) as a pandemic on 11 March 2020 , the disease remains causing enormous public health crisis globally with ongoing infections, mortality, and high economic and social impact . As of August 2021, Over 217 million people worldwide are infected with the culprit virus named severe acute respiratory syndrome-2 (SARS-CoV-2) resulting in over 4. 5 million deaths . Sadly, the COVID-19 pandemic is predicted to keep on imposing huge morbidity and mortality and disrupt societies and economies globally.
Many vaccines have been developed to control COVID-19, but people do not seem to be willing to receive them. The paper [1] reported that several factors contribute to vaccine hesitancy in covid-19. According to the WHO Strategic Advisory Group of Experts on Immunization (SAGE), vaccine hesitancy refers to a delay in accepting or refusing vaccination services despite their availability. Several organizations have identified COVID-19 misinformation, conspiracy beliefs, and social reliance as covid-19 vaccine hesitancy predictors media for COVID-19 vaccine information, gender and education levels, and so on. We based our paper on a paper written by researchers from Kenya which contains survey data collected by asking about 400 volunteers different questions to access if they intend to get vaccinated or are they hesitant to get vaccinated. Although they show some prediction they have some limitations on data-prepossessing and model evaluation. The main objective of our study is to find out the best model for predicting vaccine hesitancy of the volunteers in terms of vaccine intention. We've followed a different approach to predict results to get the best results in terms of vaccine hesitancy. Since the data was unlabelled, first of all, we took several steps to preprocess the dataset and data visualization. This part is dedicated to finding the best predictors as well as all the necessary work associated with them. Next, we build a machine learning algorithm and evaluate if the model was working properly. Finally,We got the best performance metrics at different levels of the model.

The covid-19 virus snatched a lot of lives around the world Our work had a great novelty as we built a model that predicts vaccine hesitancy and vaccine intention which helps the government or any organization to take the step against vaccine hesitancy to save themself from the covid-19 virus. No doubt, this work had a positive influence and social impact. In this research, we've used the models SVM, logistic regression, Decision tree, Randomforest and xgboost to predit the results.

II. Methodology

To create a predictive model that can help us predict whether a user is hesitant to get vaccinated, we collected the data-set that the previous research used. We used this data-set to create a classification model. Different classical machine learning classification algorithms was used to find the best model. The working sequence of the proposed COVID-19 vaccine hesitancy prediction system are represented in fig [1]

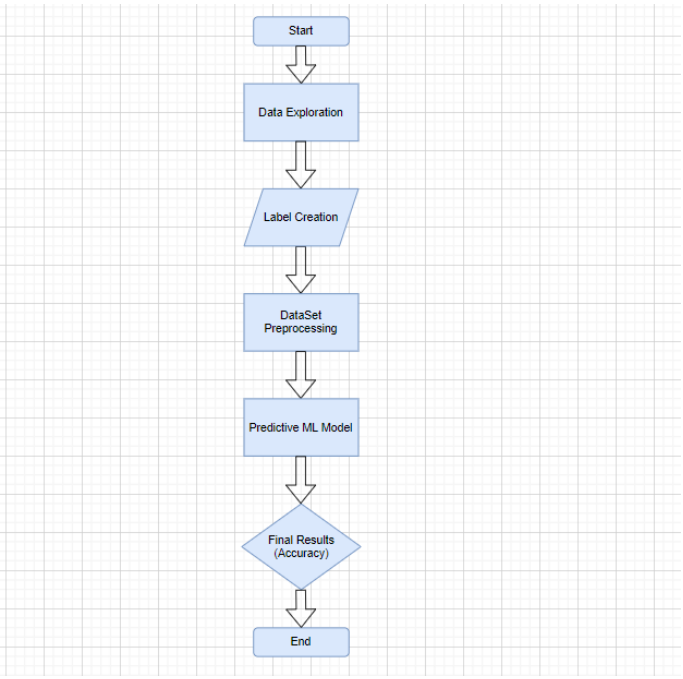

Fig. 1. Working Flowchart of COVID-19 Vaccine hesitancy Prediction

A. Data Exploration

The dataset contains information related to the CHV's opinions regarding vaccination against COVID-19. It contains 413 rows of data with 41 features and no null data. All the data is in string format. Among the different features, there was no feature that existed to be used as a label for classification. In the paper they mentioned about the CHV's vaccination intention, but this data was missing in the dataset.

B. Label Creation

Since the labels for the data are missing, now we need to figure out a way to generate the labels. Since we are specifically looking for vaccine hesitancy/vaccination intention of the volunteers, we cannot use unsupervised learning (using k-means for example) as that would create clusters and we not having any idea why each of the data is in a cluster, and with 41 features, it would be too complicated to figure that out. In the paper they mentioned some statistical information regarding what can cause vaccination intention to change. There was 6 main points that was mentioned: (1) trust in vaccine manufactures could increase vaccination intention by 25 percent, (2) trust in what the MOH says about vaccine can increase vaccination intention by 28 percent, (3) trust in vaccine safety can increase vaccination intention by 20percent , (4) if government can manage vaccine side effects, then vaccination intention can increase by 46 percent, (5) the volunteers who trust the healthcare system of their county were more likely to get vaccinated and (6) concerns about vaccine safety can reduce vaccination intention by 81 percent. These 6 values are represented in the dataset in 7 features ACQ6, ACQ7, AIQ1, AIQ4, AIQ5, AVQ1 and AVQ6, where they contain responses to questions related to the six points, and each of these features have few discrete values (we know this because the dataset was created using survey form that is available for download in the journal's page under supporting information). So, we used LabelEncoder to encode the selected features. Based on the data, we know that if we add the encoded values, we will get a higher value representing that the CHV is more likely to get vaccinated, and a lower value representing that the CHV is less likely to get vaccinated. We get an average of the seven encoded values. Now since we are looking for binary classification of whether the CHV is hesitant to get vaccinated or not, we look for a threshold value where average values higher than the threshold means not hesitant to get vaccinated or does intend to get vaccinated, and lower than threshold means the opposite. After this we get binary labels, where we get 82 percent of the users in the dataset are hesitant to get vaccinated and intends to get vaccinated, which is close to what the original paper mentioned i.e. 81 percent. Now we have labels for our unlabeled dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 413 entries, 0 to 412
Data columns (total 41 columns):
 #   Column                                                                                                                                  Non-Null Count  Dtype
---  ------                                                                                                                                  --------------  -----
 0   County                                                                                                                                  413 non-null    object
 1   District                                                                                                                                413 non-null    object
 2   Sex                                                                                                                                     413 non-null    object
 3   Age                                                                                                                                     413 non-null    object
 4   Religion                                                                                                                                413 non-null    object
 5   Education                                                                                                                               413 non-null    object
 6   Years                                                                                                                                   413 non-null    object
 7   Households                                                                                                                              413 non-null    object
 8   Have you had MoH approved training on COVID-19?                                                                                         413 non-null    object
 9   Have you been involved in educating the community on COVID-19?                                                                          413 non-null    object
 10  Your main source of Income                                                                                                              413 non-null    object
 11  Have you attended a course where COVID-19 vaccination was discussed?                                                                    413 non-null    object
 12  Do you have access to COVID-19 vaccination materials (literature, IEC materials) for quick reference when necessary?                    413 non-null    object
 13  Do you have a designated person with knowledge on COVID-19 vaccination whom you can consult when you have a question on COVID-19 vaccination?  413 non-null    object
 14  Concerning how the COVID-19 vaccines work to prevent infection                                                                          413 non-null    object
 15  To what extent do you think you are well-informed enough about COVID-19 vaccine to sensitize your family/ community members?            413 non-null    object
 16  Do you know what the COVID-19 vaccine contains?                                                                                         413 non-null    object
 17  What should the people who have received the COVID-19 vaccine do?                                                                       413 non-null    object
 18  The effectiveness of COVID-19 vaccines in preventing infection:                                                                         413 non-null    object
 19  People who have already suffered COVID-19 infection:                                                                                    413 non-null    object
 20  ACQ1: Attitude on Contextual Influences                                                                                                 413 non-null    object
 21  ACQ2: Attitude on Contextual Influences                                                                                                 413 non-null    object
 22  ACQ3: Attitude on Contextual Influences                                                                                                 413 non-null    object
 23  ACQ4: Attitude on Contextual Influences                                                                                                 413 non-null    object
 24  ACQ5: Attitude on Contextual Influences                                                                                                 413 non-null    object
 25  ACQ6: Attitude on Contextual Influences                                                                                                 413 non-null    object
 26  ACQ7: Attitude on Contextual Influences                                                                                                 413 non-null    object
 27  AIQ1: Attitude on Individual and Group Influences                                                                                       413 non-null    object
 28  AIQ12: Attitude on Individual and Group Influences                                                                                      413 non-null    object
 29  AIQ3: Attitude on Individual and Group Influences                                                                                       413 non-null    object
 30  AIQ4: Attitude on Individual and Group Influences                                                                                       413 non-null    object
 31  AIQ5: Attitude on Individual and Group Influences                                                                                       413 non-null    object
 32  AIQ6: Attitude on Individual and Group Influences                                                                                       413 non-null    object
 33  AVSQ1: Attitude on Vaccine safety and vaccination specific issues                                                                       413 non-null    object
 34  AVSQ2: Attitudeon  Vaccine safety and vaccination specific issues                                                                       413 non-null    object
 35  AVSQ3: Attitude on Vaccine safety and vaccination specific issues                                                                       413 non-null    object
 36  AVSQ4: Attitude on Vaccine safety and vaccination specific issues                                                                       413 non-null    object
 37  AVSQ5: Attitude on Vaccine safety and vaccination specific issues                                                                       413 non-null    object
 38  AVSQ6: Attitude on Vaccine safety and vaccination specific issues                                                                       413 non-null    object
 39  Source                                                                                                                                  413 non-null    object
 40  Information                                                                                                                             413 non-null    object
dtypes: object(41)
```

Fig. 2. Data-Set Features

C. Dataset Pre-Processing

For pre-processing first, we looked if there was data that would be too difficult to encode, since they are all in string format. We also looked for useless features that does not have any connection to our vaccine hesitancy label. We found 3 columns that have messy values that come from "select all that apply" kind of questions, ACQ!, ACQ2 and AVSQ5. And we found 2 columns, a source and information. All 5 columns were deleted. Then using the encoded dataset, we used pearson correlation to find features that are highly correlated and deleted one of the two that are correlated.

D. Predictive ML Model

Before we start training machine learning models using our prepared dataset, we need to create a train-test split. We need to make sure there are enough samples from each of the County in the training set as the level of education in different counties can affect vaccination intention also. We create a 60-40 train-test split to figure out the best model for classification, we tested multiple models that are intended to

be used for classification purpose. We tested SVM, Logistic Regression, Decision Tree, Random Forest, and XGBoost. For the decision tree, it was starting to overfit if we used a depth above 5, so a depth of 5 was used for the Random Forest as well as XGBoost. Fig [2] represents the total columns lists in the entire data-sets. It includes 40 different columns which are the major factors of the hesitancy

III. Results and Discussion

Finally the accuracy results we've got from the Entire Research . In fig[3] the train accuracy of SVM model is 89% and test accuracy is 82%. However, For the logistic regression the train accuracy is 87% and test accuracy is same as SVM model which is 82%. Similarly, in the random forest classification was 93% and test 85%. And lastly XGBoost had a 100% training accuracy and a 86% test accuracy. Random Forest and XGBoost had the best accuracies, but after testing multiple times, we find that XGBoost gives the best accuracy most of the time.

Model	Train Accuracy	Test Accuracy
SVM	0.8902	0.8263
Logistic Regression	0.8739	0.8263
Decision Tree	0.9552	0.8203
Random Forest	0.9308	0.8562
XGBoost	1.0	0.8622

Fig. 3. Train and test Results Accuracy.

IV. Conclusion and Future work

We have trained and tested several methods to get better results. Some of the models have given us good results, but others haven't. We obtained the best performance model from XGBoost, which gave 100% accuracy. Other models with better performance include Decison Trees (accuracy 95 percent), Random Forests (accuracy 93% percent, SVMs (accuracy 89%), and Logistic Regressions (accuracy 87%). Since we didint have own dataset we have used an open source dataset from mendeley and we worked on it. This was the limitation of this project. In future we will try to make our own dataset by applying different methods.

References

[1] COVID-19 vaccine hesitancy: Vaccination intention and attitudes of community health volunteers in Kenya.https://journals.plos.org/globalpublichealth/article?id=10.1371

[2] Khubchandani J, Sharma S, Price JH, Wiblishauser MJ, Sharma M, Webb FJ. COVID-19 vaccination hesitancy in the United States: a rapid national assessment. Journal of Community Health. 2021; 46 (2):270–7. https://doi.org/10.1007/s10900-020-00958-x PMID: 33389421

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] Qattan A, Alshareef N, Alsharqi O, Al Rahahleh N, Chirwa GC, Al-Hanawi MK. Acceptability of a COVID-19 vaccine among healthcare workers in the Kingdom of Saudi Arabia. Frontiers in Medicine. 2021; 8:83. https://doi.org/10.3389/fmed.2021.644300 PMID: 33732723

[5] Edwards B, Biddle N, Gray M, Sollis K. COVID-19 vaccine hesitancy and resistance: Correlates in a nationally representative longitudinal survey of the Australian population. PloS One. 2021; 16(3): e0248892. https://doi.org/10.1371/journal.pone.0248892 PMID: 33760836

[6] Dodd RH, Cvejic E, Bonner C, Pickles K, McCaffery KJ, Ayre J, et al. Willingness to vaccinate against COVID-19 in Australia. The Lancet Infectious Diseases. 2021; 21(3):318–9.

[7] Freeman D, Loe BS, Chadwick A, Vaccari C, Waite F, Rosebrock L, et al. COVID-19 vaccine hesitancy in the UK: the Oxford coronavirus explanations, attitudes, and narratives survey (Oceans) II. Psychological Medicine. 2020:1–15.

[8] Nzaji MK, Ngombe LK, Mwamba GN, Ndala DBB, Miema JM, Lungoyo CL, et al. Acceptability of Vaccination Against COVID-19 Among Healthcare Workers in the Democratic Republic of the Congo. Pragmatic and observational research. 2020; 11:103. https://doi.org/10.2147/POR.S271096 PMID: 33154695

[9] Fares S, Elmnyer MM, Mohamed SS, Elsayed R. COVID-19 Vaccination Perception and Attitude among Healthcare Workers in Egypt. Journal of Primary Care and Community Health. 2021; 12:21501327211013303.

[10] Government of Kenya. Kenya Community Health Policy 2020–2030. Nairobi, Kenya: 2020.

[11] Sallam M. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. Vaccines. 2021; 9(2):160. https://doi.org/10.3390/vaccines9020160 PMID: 33669441

[12] MacDonald NE. SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: definition, scope and determinants. Vaccine. 2015; 33(34):4161–4. https:doi.org/10.1016/j.vaccine.2015.04.036 PMID: 25896383