Analysis of the Social Impact of COVID Lockdown on Heart Disease

DSCC 440 Data Mining, University of Rochester

Varun Arvind and Corryn Collins

Due Date: 12/11/22

Abstract

In this study, we observe and compare the impact of the social aspect of the SARS-CoV-2 pandemic lockdown on heart disease to the cardiovascular health of Americans from 2015. The aim of this paper is to compare the heart health of two groups of people – one before the pandemic, and one during – in order to view the effects that the COVID-19 lockdown had on individual's health. A total of seven models are closely looked at when observing the possible change in cardiovascular. The model that was most consistent with the data was the logistic regression model, having the best precision and accuracy findings in relation to the other models. The results of the models show that there is not a significant impact that the pandemic had on the heart health of Americans.

Objective

The impact of the COVID-19 pandemic and lockdown had significant ramifications that compromised individuals on a global scale. On March 11, 2020, the World Health Organization declared the COVID-19 disease a global pandemic [5]. On that day, nations across the world shut their borders to international travelers, and mandated that their citizens remain in their homes in isolation, including the United States. Stay-at-home orders and social distancing was enforced by the government, which led people to fall into an unhealthy lifestyle. Americans' lives were changed: they were not moving around as much as they used to–by simply walking around in the office from meeting to meeting, or even swinging by a shopping mall. It is clear to medical researchers that the pandemic and aftermath was a global health crisis for those infected with the virus. Falling ill with the COVID virus has led to many long term side effects spanning from mental, physical, and even emotional health. We recognize these side effects, but do not wish to examine the virus itself on Americans' health. The purpose of this research is to view the *social* aspects of the government issued lock down. We examine the effect of the stay at home order placed on citizens that led their diets, movement, activity, and unhealthy life choices to decline their heart health.

Literature Review

As the pandemic surged across the world, studies have shown the negative effects of COVID–both the virus and the social phenomenon–on individual's health. One report examined the effect that COVID-19 had on heart disease and stroke mortality by examining health behaviors [8]. A survey viewed changes to lifestyle during the pandemic saw that over 27% of the participants gained weight. Americans tended to slow down during lockdown, and many did not exercise as much as they had previously. Sender et al. also looked at a study that indicated body mass indexes (BMI) in adults that were much higher than previous BMI reportings. Obesity took over during lockdown, since many people were staying home for weeks at a time and consuming a high rate of alcohol. The study references the known fact that alcohol sales and consumption significantly increased during the time of the pandemic in the US [2]. The report written by Sender et al. motivated the authors of this project to use data mining in order to view health differences brought on by COVID-19. This project intends to observe the social effects that the pandemic had on Americans cardiovascular health, with inspiration from the aforementioned article.

Methodology

For this project, we wish to look at two different datasets, heart disease data from 2015 and 2022, and various classification approaches. This project may be useful for data scientists in the medical field, as the prediction of heart disease in prone patients may reduce tragedy. This project uses various classification models for prediction. There are seven models used: Bernoulli Naive Bayes, Gaussian Naive Bayes, Random Forest, Voting, Bagging, XGBoost, and Logistic Regression.

Data Description

In order to achieve our research goal, there are two data sets used in this project: one from 2015 and the other from 2020. Both datasets were taken from the Centers for Disease Control and Prevention (CDC) collection of Behavioral Risk Factor Surveillance System (BRFSS) survey data. This annual survey is taken over the phone to probe Americans on their health, collecting data across the nation [4]. The purpose of using drastically spaced data sets from the same source is to ensure that the analyses are run on a pre–COVID time, and a time that COVID was running rampant. Both datasets were collected from the same open-access government database. The website in which these files were extracted from is a reliable and efficient source. Since the results were gathered by the same researchers, the data sets have consistent measurement attributes. There are attributes, or columns, that are seen in each data file that allow for the consistency standard to be met. These columns contain information that ranges from whether the individual has had heart disease or a heart attack in the past, if the individual is physically active or not, if the individual smokes cigarettes or not, and even touches upon education level and income. Each column has a unique benefit to the BRFSS, allowing for their collection of data to be examined with an Americans full health story painted clearly.

Data Preparation

In order to prepare the data for analysis, the datasets were both cleaned. The 2015 dataset contained more than 200 columns. In order to create a more focused analysis, only thirteen columns are chosen for the analysis seen in this report. Each column is encoded by the label that the BRFSS created, but many of these labels are unreadable for the purpose of research analysis. The labeled attribute names are replaced with more accessible names, through a renaming command. Then, the columns themselves are parsed through in order to rid the data of any unnecessary information. Occasionally, respondents did not give a “yes” or “no” answer to the surveyor, but instead either said “I don’t know” or they refused to respond to a particular question. Removal of the unanswered questions is necessary, since this process ensures that the models are utilizing essential and impactful data. In the 2020 dataset, a more cleaned dataset was found, so there was less work needed in focusing the analysis. Similar issues arose concerning the labeling of attributes, and thus the columns were also renamed to titles that better suited the pieces of data. At the end of this process, there was consistency between the two datasets. The thirteen columns analyzed include necessities such as age, and gender, but also span from if the respondent heavily consumes alcohol, their mental health, if they are diabetic, to if they have ever had a stroke.

After the data is thoroughly cleaned, it is important to understand each component of the survey and what it represents through data exploration. For example, in this report, the 'Diabetes' column has three possible answers that the respondent can give: 0 is for no diabetes or only having diabetes during pregnancy, 1 is for pre-diabetes or borderline diabetes, 2 is for yes the person has diabetes. The 'Smoker' column has also been processed to follow a Boolean method, with 0 to represent that the person does not smoke, and 1 to represent that the person does smoke. There is also an ordinal variable included by BRFSS, which is the general health attribute. This column has results that were ranked on a 1 to 5 scale, with 1 being that the individual is in excellent health, and 5 being they are in poor health. Researchers should familiarize themselves with their data, in order to better analyze models that are built with it.

Another major key that goes into model construction is to develop a test set and a training set. These two itemsets are required to be independent of each other, in order to ensure that there is no overlapping data between the sets. The training set are the samples used for model construction, with the goal of training the model with appropriate data. The testing set allows the user to see whether or not the model that was built is running correctly or not. In the attached code, the splitting of the testing and training data is seen before model building.

Exploratory Data Analysis

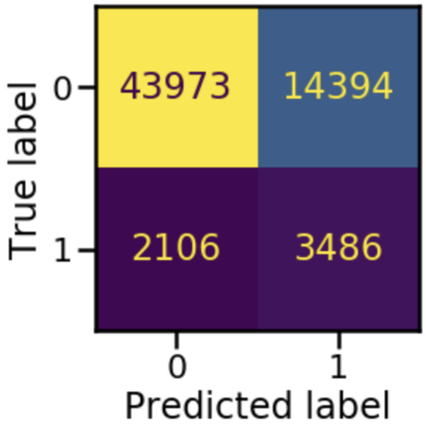
Viewing the graphs seen in the attached code, the distribution of the data can be further explored. The ages in the 2015 dataset are fairly distributed, with the majority of people over the age of 60. There also tends to appear to be an equal number of smokers between men and women, which is an interesting observation when observing heart health. The majority of people who smoke have not had a stroke, but there is a slight increase in someone being a smoker and them having a stroke than nonsmokers from 2015. The BMI concerning heart disease of individuals follows a typical curve, skewed to the right, yet the graph shows that the BMI for people who have heart disease is slightly higher than those without heart disease. The correlation graph seen in the code shows that there could be a correlation between the physical health of patients and the general health, which is expected when viewing medical profiles. There does not appear to be an obvious correlation between the other attributes of lifestyle choices, such as smoking, alcohol consumption, or BMI in the 2015 data.

Graphs for the 2020 data are very similarly distributed to the 2015 data, showing similar results to gender of the individual and their smoking status. The graph containing the information about smoking and having a stroke does not show a clear increase or decrease like the 2015 graph did. With more research, this could be reported as an improvement to stroke health. One interesting graph for the 2020 data was the physical health vs heart disease visual. The comparison can be drawn with reports of those who have heart disease. There were more people who reported that they were physically active and had heart disease than those who were not physically active with heart disease. This could be due to a bias that phone interviews have, that people overestimate their abilities when taking a survey. This is a trend that should be noted in this report. The correlation graph for the 2020 data gives less information than that of the 2015 data visual. The findings in this chart do not seem to correlate much of the attributes in this dataset, which is particularly odd when mentioning medical data. Usually, one’s life choices, such as smoking and heavy drinking, lead to their overall health depleting. These results need to be looked at with caution again, since people tend to have a more favorable self-report over the phone. This concept needs to be thought about when proceeding with the remainder of the analysis, as phone surveys may not be the most reliable form of getting medical data. Overall, the analysis can proceed, albeit with caution.

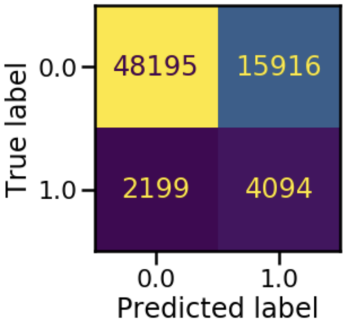
Results

Bernoulli Model

In the Naive Bayes model, the multivariate Bernoulli model is used for binary values, as seen in the heart disease data [12]. While interpreting the results of the model with the 2015 dataset, the weighted average of accuracy was 0.89, which is a somewhat high value. This number informs the researcher that the NB classifier was able to correctly predict the chance that someone has heart disease or not about 89% of the time. The model also shows other important information, such as the fact that the model was more precise when reporting those who do not have heart disease versus those who do have heart disease. In the medical sense, this information is a benefit, since the model correctly predicts true negative test results. For an issue like heart health, the worst case scenario is that a patient is diagnosed with a false negative result, or that they tested negative for heart disease, but they really have it. Overall, this model was able to put this concern to rest, meaning that it is a reliable model for the 2015 heart health dataset.

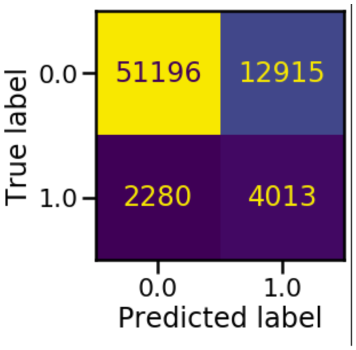


In the 2020 model, similar results were found for the Naive Bayes classifier. The weighted average of accuracy for this model was also 0.89, which was the same reporting from the 2015 dataset. With this data, the model was slightly less precise with testing whether an individual had heart disease or not. Fortunately, the model also performed well from a medical standpoint with this data. True negative tests were most accurately predicted, and false negative tests were less likely to appear than any other classification. Again, this is reassuring to those who have concerns with their cardiovascular health, ensuring that this prediction model will not allow a presence of heart disease to go unnoticed.

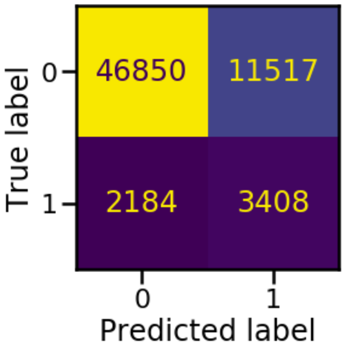


Gaussian Model

The Gaussian model is a Naive Bayes classifier, when used with a Naive Baysian algorithm, assumes a normal distribution for the data [7]. The 2015 data has a decent fit with this model, with a weighted average of accuracy of 0.89, and acceptable precision values in the classification report. The pre-COVID data has somewhat high values for the model, similar to the Bernoulli Naive Bayes model previously ran.

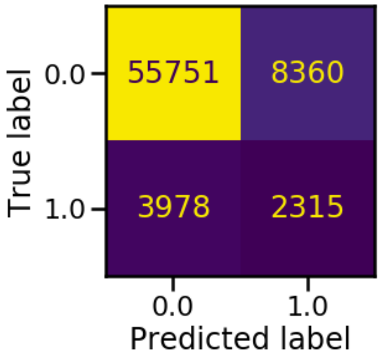


With the 2020 dataset, the Gaussian Naive Bayes model performed as well as the model did for the 2015, pre-COVID data. The values of weighted average accuracy and the precision value for a negative diagnosis were the same as the 2015 model, and the precision value for a positive diagnosis was lower by one one-hundredth. The number of false negatives, however, were much larger than seen in the 2015 model. This, again, is concerning, since the worst news for a doctor to give their patient is a false negative. The model is not a medical professional, and doctors would of course read their patients' diagnoses much closer, but in the medical field, this result could be deadly. The Gaussian NB model, with the 2020 data, was a decently fine representation and fit in this scenario, being cautious about the number of false negatives.

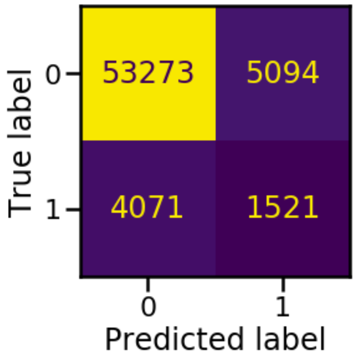


Random Forest

The Random Forest is an ensemble classifying model that utilizes many decision trees. This model aims to form a more accurate prediction of results by using a combination of the trees [9]. For the 2015 dataset, the weighted average of accuracy was 0.87, which is a fair numerical representation. This number shows that the Random Forest classifier was able to correctly predict the chance that someone has heart disease or not about 87% of the time. The model did not perform too well when recognizing false negatives. This was one of the higher classified groups, which is concerning for individuals that may have tested negative for heart disease, but are actually carrying the disease in them. When using this model with the 2015 heart health dataset, researchers should be cautious of its classifying results, since there tended to be a concerning amount of false negatives.

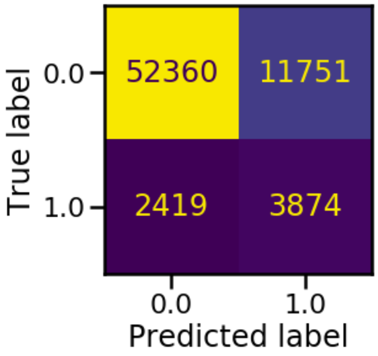


In the 2020 model, similar results were present in the Random Forest classifier. The weighted average of accuracy for this model was also 0.87, which was the same reporting from the 2015 dataset. With this data, the model surprisingly had the same precision with testing whether an individual had heart disease or not as was seen in the data from 2015. Unfortunately, the model also did not perform well from a medical standpoint with this data. True negative tests were once again the most accurately predicted, but the false negative tests were also more prevalent than true positives. Again, this is a concern to those who actually do have cardiovascular health problems, notifying researchers that this prediction model may not be the most optimal classifier to be used with this data.

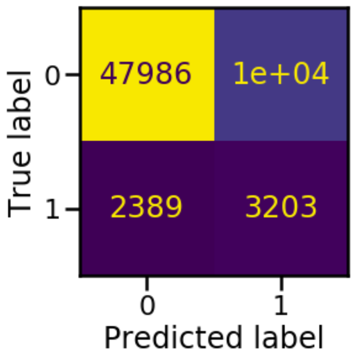


Voting Classifier

The Voting classifying model is used as a classification method to group many different machine learning models [11]. In this model, the Voting model utilizes the three aforementioned models: Bernoulli, Gaussian, and Random Forest. The results from the Voting model are consistent with the findings mentioned above. For the 2015 data set, the weighted average of accuracy of Voting is 0.89, which is a good accuracy score when combining the models. The Voting model also is true to the results of the Bernoulli classifier for 2015, as there were the least amount of false negatives reported out of all classifying categories. This model does have the highest precision value for the positive heart disease tests out of the three others, sitting at 0.25. This is a very subtle yet very reassuring increase. The Voting model performed well on the pre-COVID dataset taken from 2015.

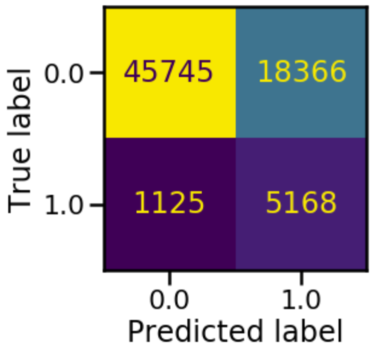


Observing the Voting classifier for the 2020 model, the classification report and confusion matrix seen in the code has a very similar output to the data collected in 2015. With the same weighted average of accuracy as above, 0.89, this model fits the 2020 data well. The precision metrics are also consistent in this model, with only one one-hundredth. To data science researchers and medical staff, this is a very good sign to note that the model is compatible with the data and the information given. It is definitely reassuring to see that the false negatives continue to be the least frequent classification in the model. Again, this is a good model that has fit the 2020 data in an accurate and acceptable way.

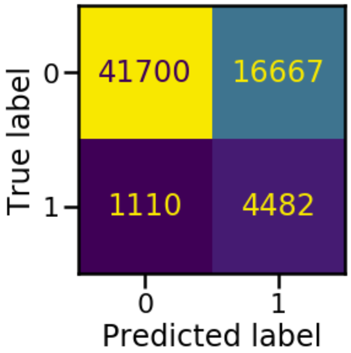


Bagging Classifier

The Bagging classifying model is used with the decision tree classifying model. This is a model that averages various models to prevent the overfitting of data [10]. In 2015, the Bagging model performed extremely well. The weighted average of accuracy for the Bagging model was 0.91, which is the highest accuracy average that has been reported thus far. This number informs the researcher that the Bagging classifier was able to correctly predict the chance that someone either has heart disease or not about 91% of the time. This is only a slight step up from the previous accuracy reporting of 0.89, yet to data scientists, this reporting between 0.90 and 1.0 is extremely reassuring when using the model. The precision of predicting a negative test was also 0.98. This model had the least number of false negatives by far. Overall, this model performed extremely well with the 2015 dataset.

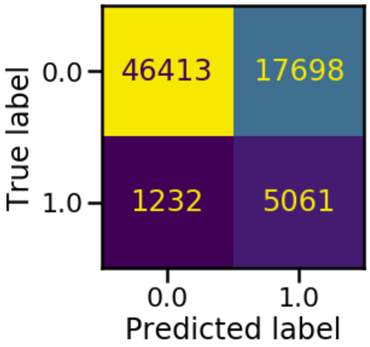


The Bagging model performed equally as well with the 2020 model. The weighted average accuracy of the model was the same as reported above, 0.91. The precision score was only slightly lower than the 2015 dataset, but had equally as good values in the output. This also is emphasized with the great finding that the false negative heart disease test results made up a very small percentage of the model. The Bagging model performed well with the pandemic of 2020's data, as well as the 2015 data set.

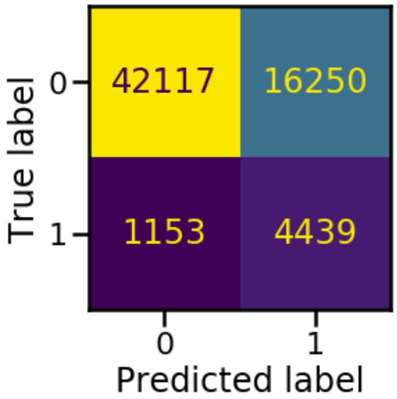


XGBoost Classifier

XGBoost is a decision tree algorithm that uses boosting to help improve performance [1]. On the 2015 data, the XGBoost model performed very well, with a weighted average of accuracy of 0.91. The precision of negative and positive heart detection tests also consistently performed great, with a precision of 0.97 and 0.22, respectively. This value is in line with the great performing training and testing sets.

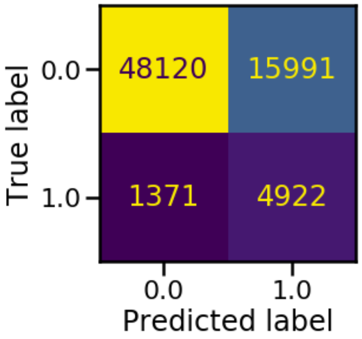


On the 2020 data, the XGBoost model performed equally well. The weighted average of accuracy as well as the precision values were consistent with the data from 2015. Again, this model is a great tool when observing both datasets from the BRFSS.

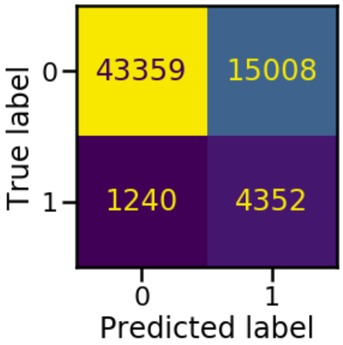


Logistic Regression

The logistic regression model is used when analyzing the probability of an event occurring through a linear combination and regression model. It is mainly used to show relationships that are present in the given data [3]. The logistic regression model on the 2015 data provided a great analysis on how the data worked with the model. The weighted average of accuracy was 0.91. There were barely any false negatives with the model, which shows the quality of the model.



The data from 2020 proved to have similar results as the 2015 model. The weighted average of accuracy and the precision concerning negative diagnoses was the same, and the precision of positive heart disease diagnoses was less than that of the 2015 model by two one-hundredths. The logistic regression model is encouraged to be examined when observing the overall fit of all the models against each other, in the discussion section.



Conclusion

All of the models fit well with both datasets. Overall, the model that worked the best with the 2015 data was the logistic regression model. This model had the highest values for weighted average of accuracy as well as precision for positive and negative diagnoses. This model also had the most optimal false negative results out of the seven models examined. It is suggested that this model is used for analysis of the BRFSS datasets. The worst performing model for the 2015 pre-COVID dataset is Random Forest model. This model consistently had the worst results with the data, in terms of precision and accuracy values reported. This model also had the largest number of false negatives in the confusion matrix, which is an extreme red flag in medical research. The Random Forest model is not recommended to be used with the dataset.

Along with the results from the 2015 dataset, the 2020 dataset taken during lockdown performed only slightly worse with the seven models. The best performing model with the data taken in 2020 was also the Bagging model. With this data, the linear regression model and the Bagging model were very closely well suited, but the results of the Bagging model persisted ever so slightly. The values show great performance, but were just below those seen in the 2015 dataset models. The weighted average of accuracy, 0.91, and the precision values: 0.97 for negative diagnosis and 0.21 for a positive diagnosis, have been seen to be consistent with the values seen in other models. The piece that made the Bagging method unique was the very low number of false negatives. This is the lowest reported number for false positives in the report, which is very significant. This proves that the Bagging model should be used with the 2020 dataset. On the other hand, the worst performing model with the during-pandemic data is again the Random Forest model. This model had a low accuracy report, and a very concerning number of false positives, similar with the findings of the 2015 dataset. The confusion matrix showed an extremely high value for false negatives, and a low number of true positives. This issue is absolutely worrying for this research, and this model is not recommended to be used with the 2020 heart dataset.

Overall, the models performed well with both datasets. After reviewing each machine learning model on the data, the goal of this paper can be discussed. The 2015 dataset performed slightly better than the 2020 dataset with every model. This could be due to differences in the reporting of each dataset. For example, individuals may have responded on the phone, reporting less health problems in 2015. During the pandemic, people could have responded in a more negative way, due to a possible mental health decline of the lockdown. Physical health during the pandemic could have also been worse than in 2015, inferring that the average American was more unhealthy in 2020. The precision measurement between the two datasets for each model is also consistently lower for the 2020 dataset. The pandemic was a tough time in Americans lives, which could indicate results that are atypical, or even unprecedented. Due to symptoms of COVID, both medically and socially, people could have been reporting data that was not as correlated as the data from 2015. The models may have noticed the fact that the same trends were not being followed in 2020. For example, COVID may have been the motivator that declined someone's physical activity levels, not heart disease. This is one scenario of a possibility during 2020.

Comparing the 2015 dataset with the 2020 dataset, heart disease did not seem to get worse due to the pandemic. If anything, the results of heart disease seemed to stay pretty consistent over the five years. There is little to no evidence with this data that suggests that heart disease became worse as a result of the pandemic. To researchers, these results are not ideal, but mining data related to this problem will continue, and may produce a different result. To reiterate, caution needs to be taken when researching phone surveys, but this has been taken into consideration for this report and is not believed to impact the results.

The limitations of this study include the range of the dataset collections. In the grand scheme of things, five years is not an extensible amount of time. For this research, the five years between the collection of the 2015 dataset and the 2020 dataset could be influential to the results. To ensure that there were not skewed health results during the time of the COVID virus, the choice to select a dataset from long before the illness was around. Comparing the time in the world from 2015 to 2020 may have led to an overall decline of Americans' health, which may not be attributed to COVID. This is also something that the report recognizes as a possibility, along with any number of attributes that could have led to a shift in health in the five years. With more research and resources, the data and the analysis could be improved with a more medical outlook.

References

[1] B. T, “Beginner's Guide to xgboost for classification problems,” *Medium*, 09-Dec-2022. [Online]. Available: https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390. [Accessed: 11-Dec-2022].

[2] Brown, “Alcohol consumption during the COVID-19 pandemic projected to cause more liver disease and deaths,” *Massachusetts General Hospital*. [Online]. Available: https://www.massgeneral.org/news/press-release/alcohol-consumption-during-the-covid-19-pandemic-projected-to-cause-more-liver-disease-and-deaths. [Accessed: 11-Dec-2022].

[3] C. Subasi, “Logistic regression classifier,” *Towards Data Science*, 04-Mar-2019. [Online]. Available: https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9. [Accessed: 30-Oct-2022].

[4] “CDC - about BRFSS,” *Centers for Disease Control and Prevention*, 16-May-2014. [Online]. Available: https://www.cdc.gov/brfss/about/index.htm. [Accessed: 11-Dec-2022].

[5] D. Cucinotta and M. Vanelli, “WHO Declares COVID-19 a Pandemic,” *Acta Biomedica Atenei Parmensis*, 2020. [Online]. Available: https://doi.org/10.23750/abm.v91i1.9397. [Accessed: 11-Dec-2022].

[6] “Heart disease facts,” *Centers for Disease Control and Prevention*, 14-Oct-2022. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed: 29-Oct-2022].

[7] R. Gandhi, “Naive Bayes Classifier,” *Towards Data Science*, 05-May-2018. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c. [Accessed: 29-Oct-2022].

[8] S. Sender, R. S. Blumenthal, G. Sharma, and P. Kohli, “Covid-19's impact on heart disease and stroke mortality,” *American College of Cardiology*, 25-May-2022. [Online]. Available: https://www.acc.org/Latest-in-Cardiology/Articles/2022/05/25/12/10/COVID-19s-Impact-on-Heart-Disease-and-Stroke-Mortality. [Accessed: 29-Oct-2022].

[9] “Sklearn Random Forest Classifier,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: 29-Oct-2022].

[10] “Sklearn.ensemble.BaggingClassifier,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html. [Accessed: 11-Dec-2022].

[11] “Sklearn.ensemble.VotingClassifier,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html. [Accessed: 11-Dec-2022].

[12] “Sklearn.naive_bayes.Bernoullinb,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html. [Accessed: 11-Dec-2022].

[13] “Traditional risk factors predict heart disease about as well as sophisticated genetic test, study suggests,” *UT Southwestern Medical Center*, 18-Feb-2020. [Online]. Available: https://www.utsouthwestern.edu/newsroom/articles/year-2020/predicting-heart-disease.html. [Accessed: 29-Oct-2022].