HÖGSKOLAN DALARNA

Statistical Learning Home Exercise

Exercise 1 – Exploring American Beliefs in the COVID-19 Conspiracy Theory – Coronavirus is a Chinese Bioweapon?

Review by Saumya Gupta

h20saugu@du.se

April 26, 2021

Introduction

This report aims to describe the author's solutions to the first home exercise of the Statistical Learning course at Dalarna University during the second period of the Masters in the Data Science programme. The report will explain the two problem statements to be solved as part of the exercise, explaining the provided dataset. After this, the report will state the data manipulations, describe the methods and models used, establish the results, and finally discuss the findings for each of the two solutions. The author will also list the limitations and violations of model assumptions in the approach.

Background: The COVID-19 pandemic has had many conspiracy theories surrounding the origin of the virus. One of which says that the Chinese government developed the virus as a bioweapon. Previously, to answer questions related to exploring American beliefs in this conspiracy, a survey was conducted in the year 2020 by Stecula and Pickup in their research. The author will use the same data set to provide solutions to two tasks as instructed by the home exercise 1:

1. Find out if any individual type of American believes in the mentioned conspiracy more than the other type.

2. Build a suitable prediction model to predict any specific American's degree of belief in the conspiracy theory.

Data and Related Methods

Dataset: As mentioned previously, the data comes from an original survey of 1009 US adults and reflects the US population. The survey measured the respondents' degree of belief in the conspiracy theory by their agreement to the statement – The Chinese Government developed the coronavirus as a bioweapon (Pickup & Stecula, 2021). To facilitate their research, Stecula and Pickup ranked several attributes and behaviours of the respondents (30 features), mainly on a scale of 1 to 4. For example, how frequent viewers of conservative news are they? How strong are their populist views? How much do they trust in the CDC? Their partisanship? Along with these, it recorded other demographic information such as age, sex, gender, and ethnicity. Table 1 gives a detailed explanation of all variables in the dataset and their usage in the analysis.

Removing Unnecessary Variables: The following variables were removed/ignored in the analysis:

- cov_beh_sum – variable measures the effect of the response – the degree of belief in COVID conspiracy. According to the author, the analysis should not consider it for the position of a predictor. A predictor fundamentally influences the response and is not the effect of it.

- cons_covax & cons_covax_dummy – variable measures the degree of belief in another conspiracy, that there is a coronavirus vaccine that the national governments and pharmaceuticals companies will not release. Though there is a significant correlation between the response and this variable ($\rho = 0.6$), the analysis does not include this since there could be independent variables in the model, which influence both the conspiracies, for example, strong populist views. If so, we need to suppress the confounders' effect to evaluate this variable's effect on the response, which is not the goal of this analysis.

- pid3 – variable means the same as pid2, where pid2 seems to be a dichotomised version of pid3. For easy understanding, this analysis uses pid2.

- weight – variable can be ignored as per the problem statement. It means that this statistical procedure will weigh each case/ individual/ data point equally.

Table 1. Variables' Type and Usage Description

Variables	Form in Raw Data	Considered for Analysis	Form of Usage in Analysis	Usage Details
trust_1	Discrete Numerical	Yes	Numeric	
populism_1	Discrete Numerical	Yes	Numeric	
populism_2	Discrete Numerical	Yes	Numeric	
populism_3	Discrete Numerical	Yes	Numeric	
populism_4	Discrete Numerical	Yes	Numeric	
populism_5	Discrete Numerical	Yes	Numeric	
age	Continuous Numerical	Yes	Numeric	
gender	Discrete Numerical	Yes	Categorical	Problem 1: Factored Problem 2: Factored and Dummy coded (because of KNN)
hhi	Discrete Numerical	Yes	Numeric	
hispanic	Discrete Numerical	Yes	Numeric	
cov_beh_sum	Discrete Numerical	No	Numeric	
cons_biowpn	Discrete Numerical	Problem 1: No Problem 2: Yes	Categorical	Factored
cons_covax	Discrete Numerical	No		
cons_biowpn_dummy	Discrete Numerical	Problem 1: Yes Problem 2: No	Categorical	Factored
cons_covax_dummy	Discrete Numerical	No		
white	Discrete Numerical	Yes	Numeric	
highered	Discrete Numerical	Yes	Numeric	
idlg	Discrete Numerical	Yes	Categorical	Problem 1: Factored Problem 2: Factored and Dummy coded (because of KNN) Dichotomised into Conservative (Extremely conservative, Conservative, Slightly conservative) and Liberal (Moderate, Liberal, Slightly liberal, Extremely liberal)
pid3	Discrete Numerical	No		
pid2	Discrete Numerical	Yes	Categorical	Problem 1: Factored Problem 2: Factored and Dummy coded (because of KNN)
md_radio	Discrete Numerical	Yes	Numeric	
md_national	Discrete Numerical	Yes	Numeric	
md_broadcast	Discrete Numerical	Yes	Numeric	
md_localtv	Discrete Numerical	Yes	Numeric	
md_localpap	Discrete Numerical	Yes	Numeric	
md_fox	Discrete Numerical	Yes	Numeric	
md_con	Discrete Numerical	Yes	Numeric	
md_agg	Discrete Numerical	Yes	Numeric	
weight	Continuous Numerical	No		
rw_news	Continuous Numerical	Yes	Numeric	Average of conservative news – md_radio, md_national md_broadcast, md_localtv md_localpap and md_agg
ms_news	Continuous Numerical	Yes	Numeric	Average of mainstream news – md_fox and md_con
populism123	-	Yes	Numeric	Created as the average of populism_1, populism_2 and populism_3
populism45	-	Yes	Numeric	Created as the average of populism_4 and populism_5

Dealing Missing Data: The dataset consists of missing values in 173 rows after removing the variables mentioned above. Since all the variables left could seem relevant for both inference and prediction, we cannot remove any variables further, neither can we impute data values for a survey. So, we follow the 'usual best' approach of omitting these 173 rows (17.14%) from our dataset.

Categorical Variables Usage: The dataset consists of many numerically coded ordinal (for example, trust_1) and nominal (for example, gender) variables. We factor the nominal variables and either dummy code them explicitly (Problem 2) or use R's reference level coding (Problem 1). At the same time, we treat the ordinal variables as numeric to preserve the information in the ordering. This approach requires the assumption that the numerical distance between each set of successive categories is equal (Grace-Martin, 2018). For our case, this assumption holds, and hence our analysis could render results close to reality. The variable, idlg, was dichotomised into Liberal and Conservative, following those who are not Conservative in any way are Liberal. See Table 1 for details.

Solution 1

Model for Inference – Logistic Regression

Since the first problem asks of a model is mere inferences, we use logistic regression, which is suitable for making inferential decisions but not that good of a prediction classification model.

Model Usage Analysis:

- It makes no assumptions related to the distribution of predictors within the classes.
- It will show how effective a predictor is (coefficient size) and the direction of association (positive or negative).
- We can use it for two-class response variables, which is suitable since the exercise instructs to use the two-class variable – cons_biowpn_dummy, as the response.
- It requires no multicollinearity between independent variables. Table 2 shows no multicollinearity with the help of the Variance Inflation Factor score, as all values lie close to 1.

Note: It is important to note that we are not using logistic regression for prediction, and hence, we will not evaluate it based on how good of a fit it provides. Therefore, we will not split the data into training and test set or evaluate the model with the test.

Odds Ratio: Odds of success refers to the ratio of the probability of success and the probability of failure. Success for a respondent here is the respondent believing in the conspiracy theory, which is event C's occurrence.

Event C: Respondent believing in the conspiracy theory.

Event A_1: American a_1 is the respondent.

Event $A_{2:}$ American a_2 is the respondent.

$$Odds\ (success\ a_1)\ =\ \frac{P\ (\ C\ |\ A_1\ )}{1 -\ \ P(\ C\ |A_1\ )}$$

The odds ratio refers to the ratio of odds of success for a_1 and odds of success for a_2.

$$Odds\ Ratio\ = \left.\frac{P\ (\ C\ |\ A_1\ )}{1 -\ P(\ C\ |A_1\ )}\middle/\frac{P\ (\ C\ |\ A_2\ )}{1 -\ P(\ C\ |A_2\ )}\right.$$

In the logistic regression model, for a particular set of input variables $Xp\ = \{X_1, X_2, X_3 \ldots X_k \ldots X_n\}$ influencing the response, the odds of success of response change by a factor of e^{β_k} for a unit change in the input variable X_k keeping all the other $n-1$ input variables constant where e is the Euler's constant and β_k. In other words, for two individuals a_1 and a_2 with the values of x and $x+1$, respectively, for a predictor variable X_k, let us say age, with the coefficient estimate of β_k calculated by the logistic regression model, the following holds, keep all other predictor variables constant:

$$Odds\ (success\ a_1) *\ e^{\beta_k}\ = Odds\ (success\ a_2)$$

It will also hold for dummy coded categorical variables, where instead of x to $x+1$, we move from 0 to 1.

Results

Variables ms_news and rw_news are aggregates of various sub-components as listed in Table 1. We can also see the high correlations between them in Fig. 1. They represent aliased variables in the model with their sub-component covariates. Hence the model can only accommodate any one of the two. Accordingly, keeping all the one-factor covariates in the model (excluding the combined covariates), we get the estimates shown in Table 2 along with their significance level (shown with *). We notice that many of the variables show statistically insignificant relationships with the response in the model using all predictors. In contrast, the degree of agreement to populist statements 1, 3 and 5, partisanship and frequency of watching conservative news – Fox News reject the null hypothesis (p-value < 5%) and show their importance as significant predictors of the response.

Table 2. Summary of Model with All One-Factor Terms (AIC = 975.33)

	Estimate	Std. Error	Pr (>|z|)	VIF
(Intercept)	-3.9625501	0.6634568	2.34e-09 ***	-
md_agg	0.0916115	0.0923831	0.32137	1.719098
md_localpap	-0.0514423	0.0906248	-0.57028	1.666559
md_localtv	-0.0213835	0.1143984	0.85172	1.980056
md_broadcast	0.1214492	0.1101028	0.27000	2.121957
md_national	-0.0549447	0.0997623	0.58180	1.899818
md_radio	-0.0821949	0.0892462	0.35706	1.536962
md_con	0.1932418	0.1011793	0.05615 .	1.936002
md_fox	0.3624776	0.0893881	5.01e-05 ***	1.676965
pid2RepL	0.6064028	0.2116970	0.00418 **	1.752750
populism_1	0.5185043	0.1182782	1.17e-05 ***	1.330868
populism_2	-0.0375026	0.1151814	0.74473	1.314105
populism_3	0.2195390	0.0956531	0.02172 *	1.197327
populism_4	0.1114637	0.1091595	0.30720	1.414649
populism_5	0.4897791	0.0947331	2.34e-07 ***	1.293405
white	-0.3348308	0.2062570	0.10451	1.349814
highered	-0.0005385	0.1732696	0.99752	1.145001
hispanic	-0.0366334	0.2555501	0.88601	1.113333
genderMale	0.1858071	0.1643116	0.25813	1.060217
age	-0.0110236	0.0057523	0.05532 .	1.451355
hhi	-0.0112445	0.0134320	0.40251	1.164875
trust_1	-0.1243246	0.1121958	0.26782	1.191160
idlgLiberal	-0.0893416	0.2028969	0.65970	1.539075

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

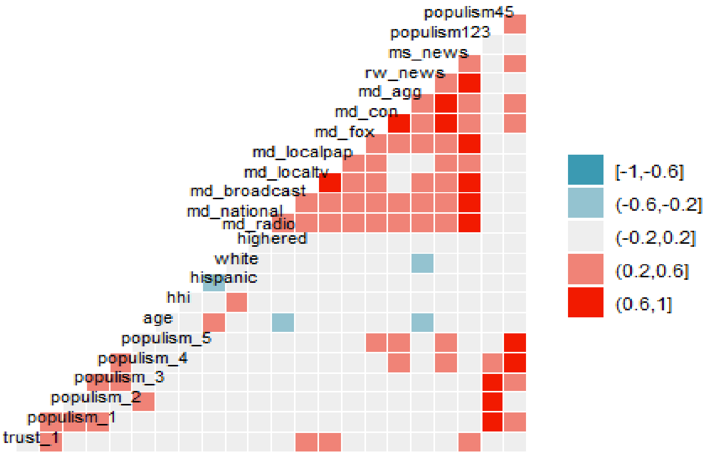

Fig. 1. Correlation plot for all numeric variables in the conspiracy dataset.

We try to make the model better for better inferences by subset selection with the AIC (Akaike Information Criterion) as the criteria and remove the variables with statistically insignificant influence. We see from Fig. 1 since the correlation between md_fox and md_con and similarly between md_broadcast & md_localtv are somewhat high, we go with the aggregates ms_news and rw_news containing the measures of all sub-components. Since the correlation between populism_4 and populism_5 is not high enough, we take these separately, as shown in Table 3.

Table 3. Summary of Model after Variable Selection and best AIC (955.36)

	Estimate	Odds Ratio	Std. Error	Pr (>|z|)	VIF
(Intercept)	-4.171100	-	0.508134	2.24e-16 ***	-
pid2RepL	0.665644	1.945743	0.179603	0.00021 ***	1.277990
rw_news	0.550093	1.733414	0.083506	4.47e-11 ***	1.090021
populism_1	0.486770	1.627052	0.108367	7.06e-06 ***	1.130246
populism_3	0.231704	1.260746	0.091833	0.01163 *	1.122234
populism_5	0.534745	1.707013	0.084585	2.58e-10 ***	1.040175
white	-0.387267	0.6789098	0.195177	0.04723 *	1.223871
age	-0.010827	0.9892314	0.005068	0.03264 *	1.141533

We further verify this by building a more reduced model with populism123 and populism45 variables in place of subcomponent covariates created based on the survey question types shown in Table 1. We can see the summary in Table 4.

Table 4. Summary of Model after Variable Selection with Aggregate Terms (AIC = 966.21)

	Estimate	Std. Error	Pr (>|z|)	VIF
(Intercept)	-4.061361	0.530439	1.91e-14 ***	-
pid2RepL	0.616730	0.175543	0.000443 ***	1.242002
rw_news	0.525644	0.082436	1.81e-10 ***	1.082575
populism123	0.589023	0.140469	2.75e-05 ***	1.112607
populism45	0.686680	0.106598	1.18e-10 ***	1.076129
white	-0.419517	0.193975	0.030562 *	1.228233
age	-0.014109	0.005013	0.004890 **	1.133935

Though there is not a very significant difference in AICs of models in Table. 2, 3 and 4, we will choose Table 3 coefficient estimates with the best AIC for calculating inferences. Table 3 also shows the calculated odds ratios for each of these significant predictors for 1 unit increment, further plotted in Fig. 2. We infer that:

- The odds of an American with more robust populist views, as seen through populist question 1 (say, four on a scale of 1 - 4) in believing that the Chinese Govt. developed COVID-19 as a bioweapon is 1.6 times that of a person with lesser strong populist views (say, three on a scale of 1 - 4). Alternatively, a person who strongly agrees to the populist statement 1 has 330.7 % more odds of believing the conspiracy than a person who strongly disagrees. Similar stand for populist question 5. For populist question 3, the percentage is not this great, but still very high, making these populist questions significant influencers of the response. Of course, these calculations are with keeping other predictors constant.
- An American identifying as a Republican (including leaners) has 94.6 % more such odds than identifying as a democrat (including leaners). It is the most influencing variable according to our analysis.
- An American who frequently watches conservative news has 420.8 % more odds of believing in the conspiracy than the one who watches no conservative news.
- Race (white or not) and age also are significant negative influencers, as shown in Fig. 2. A white American has 32 % fewer odds than a non-white person. It is only a 1.1 % decrease in odds with 1 unit of increase in age.

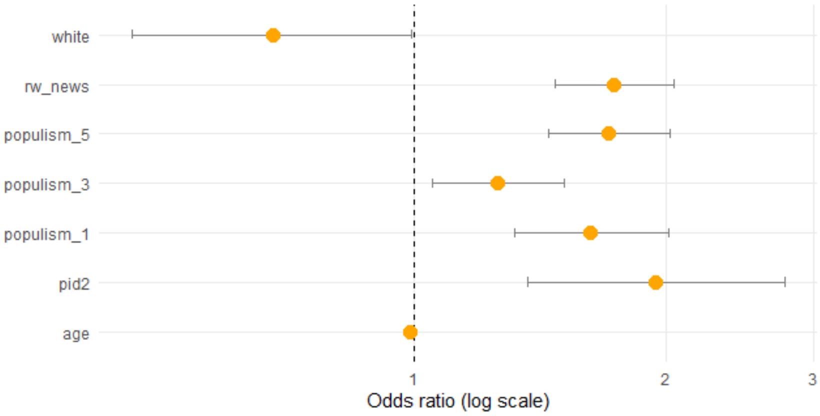

Fig. 2. Odds ratios with 95% Confidence Interval (CI) on a logarithmic scale for the chosen model.

Discussion

This analysis shows that while many variables show an insignificant relationship with the degree of belief in the conspiracy, for example, trust in CDC, ideology or higher education, there were variables that are pretty significant influencers to the response and lead us to conclude: Americans with populist views / more substantial populist views, who identify themselves as Republicans or are leaning towards it, who watches / more frequently watches conservative news channels like Fox News, have relatively lesser age, and lastly, non-whites have significantly more percentages of odds as compared to others. Of course, these are when keeping all other variables constant (for example, all seven other variables constant while considering populism_1 in Table 3). To consider all the variables together, we need to calculate the probabilities for each case to compare.

Solution 2

Model for Prediction – Naïve Bayes wins over LDA, QDA and KNN!

Since what the second problem asks of a model is predicting the degrees of belief of an American in the conspiracy theory (cons_biowpn), which is a multiclass variable, we analyse the performance and reliability of 4 multiclass classification models: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K- Nearest Neighbours (KNN) and lastly, Naïve Bayes. Based on their performances, we propose one of them as the solution. Here, we want to measure the performance of our models based on various evaluations on a test set, and hence, we will split our data into training (80%) and test (20%) set. While doing that, we will look at various single label multiclass evaluation metrics – accuracy, balanced/weighted accuracy, log loss and area under the curve (AUC).

Model Usage Analysis and Preliminary Tests:

1. **LDA:**
   - It assumes Gaussian distribution for predictor variables $Xp = \{X_1, X_2, X_3 \dots X_k \dots X_n\}$ for every class k of response. To check if several variables are normally distributed as a group, perform the energy test for multivariate normality. The test shows strong rejection of the null hypothesis, and hence we can conclude that variables do not follow multivariate normal distribution for all four classes.
   - It assumes homogenous variance-covariance matrices across all classes of response. To check this assumption, we use Box's M-test. It is worth noting that the Box M test is sensitive and can detect even minor departures from homogeneity (Schmidt, 2018). Test show that we have a problem of heterogeneity of variance-covariance matrices.
2. **QDA:**
   - It can work great when there are many training observations, and the assumptions of LDA for normality holds, but the equal variance is not realistic.
   - In this analysis, training data consists of 668 observations, which the author does not regard as considerably large. The absence of normality is a problem in our case.
3. **Naïve Bayes:**
   - It needs less training data and can perform better in the case of independent predictors. Of course, we cannot have completely independent predictors in real life.
   - It performs well in the case of categorical variables compared to numerical. Our data is mixed and has more numerical input variables.
   - It assumes normal distribution for numerical variables. As mentioned, this is not reliable for our dataset.
4. **KNN:** It does not involve any distribution-related assumptions, but there is a question of "which k to use". Our analysis uses the k value using 10-fold cross-validation (CV).

Preliminary analysis tells that since our predictors do not show normality, KNN should perform the best if not well. We will check the performance of all the models.

Variable Selection: With the backward selection, variable selection results vary based on the classification function used and the performance measure for selection. Instead of making the variable selection using each classification method, we use all the one-factor predictor terms for all models. As mentioned before, we use ms_news and rw_news in place of all the sub-component covariates due to collinearity issues. For the reasons mentioned before, we use populism variables separately instead of looking at combined effects through populism123 and populism45.

Data Manipulation: All the preliminary data manipulations and assumptions of variable types remain the same as for problem 1, but here, we need to create dummy columns for gender, idlg, and pid2 explicitly because of the inability to work with KNN. In problem 1, we find a better logistic regression model with variable selection, but here we do not do that.

Evaluation Metrics:

1. **Accuracy:** The multiclass accuracy is the average number of correct predictions. The higher, the better.
$$Accuracy = \frac{Total\ correct\ predictions}{Total\ instances}$$
2. **Weighted Accuracy:** The multiclass weighted accuracy calculates the accuracy by assigning weight to the number of correct predictions in each class based on the number of observations belonging to that class. For our calculations, we take the weight $w_k = \frac{1}{G_k}$ for class k, where G_k is the number of observations belonging to the class k. After taking the weighted sum, we divide by the number of classes. If K denote the total number of classes and r_k denotes the number of correct predictions in class, k, then the equation below shows the calculation of weighted accuracy. The higher, the better.
$$Weighted\ Accuracy = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{G_k} r_k$$
3. **Log Loss:** Consider that N denotes the total number of observations. The total number of classes is K. k is the actual class of the data point n. P_{nk} is the predicted probability of observation n belonging to class k. The equation below gives the calculation for multiclass log loss. It is a loss; higher is not good.
$$Log\ Loss = \frac{1}{N}\sum_{n=1}^{N} \log P_{nk}$$
4. **AUC:** For this, we use multiclass.roc function from pROC (Robin, et al., 2011) computes the multiclass AUC defined by Hand and Till. The higher, the better.

Results

Table 5 shows the results for models that we want to analyse based on the evaluation metrics mentioned before.

Table 5. Evaluation on Test for Naïve Bayes, LDA, QDA and KNN

Model	Accuracy	Weighted Accuracy	Log Loss	AUC
Naïve Bayes	0.4821429	0.4873505	1.403069	0.7089
LDA	0.4345238	0.4376968	1.272546	0.6947
QDA	0.3809524	0.3825466	1.740215	0.6561
KNN ($k = 8$)	0.2678571	0.2648089	0.1910587	0.5601

Looking at the accuracy and weighted accuracy, none of the models performs well with the test set. Nevertheless, comparing, we can say Naïve Bayes works the best of the four if we look at the accuracy, but not when we look at log loss. Log loss takes the predicted probability of the actual class by the model into consideration, whereas accuracy takes no. of correct predictions. If we add the AUC result to our consideration, we can say that the Naïve Bayes classifier works the best. On the other hand, KNN (calculated through CV) performs the best in log loss but the worst in all other metrics, with an AUC of just 0.5601, slightly better than the worst model.

Discussion

The analysis tests the LDA, QDA, KNN and Naïve Bayes models on the new data for 168 Americans. Our results show that Naïve Bayes performs the best, though it cannot be considered a good estimator, given the low accuracy, rather a lousy estimator. We expected LDA and QDA to not perform well because of the violation of assumptions. KNN, which does not assume any distributions, was expected to work better if not best, but KNN (with 10-fold cross-validation k) has the worst evaluation. Similar values of accuracy and weighted accuracy for a model are due to the balanced data for classes in the test and the training set. Based on our analysis, we choose the Naïve Bayes Model.

Limitations

The author identifies the following limitations to the analysis:

1. Though the missing data represents 0.88% of the total data, the author has removed the corresponding rows in the data frame, representing 17% of the total rows available initially. We could have possibly lost some information affecting the statistical power of our analysis.
2. We dichotomised the idlg variable and used the dichotomised version pid2 instead of pid3 for ease of inference and in the interest of time. The practice, in general, has substantial negative consequences, as listed and proven by the study by (MacCallum et al., 2002), which motivates to reduce the practice of dichotomisation and its negative impact.
3. The study did not go into interaction terms and other transformations, and the author omitted this in the interest of time.

References

Grace-Martin, K. (2018). *Pros and Cons of Treating Ordinal Variables as Nominal or Continuous*. Retrieved April 26, 2021, from The Analysis Factor: https://www.theanalysisfactor.com/pros-and-cons-of-treating-ordinal-variables-as-nominal-or-continuous/

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the Practice of Dichotomisation of Quantitative Variables. *Psychological Methods, 7*(1), 19-40. doi:DOI: 10.1037//1082-989X.7.1.19

Pickup, M., & Stecula, D. A. (2021). How populism and conservative media fuel conspiracy beliefs about COVID-19 and what it means for COVID-19 behaviours. *Research and Politics, 8*(1). doi:10.1177/2053168021993979

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyse and compare ROC curves. *BMC Bioinformatics, 12*, 77.

Schmidt, P. (2018, February 19). *Assumption Checking of LDA vs QDA – R Tutorial (Pima Indians Data Set)*. Retrieved April 26, 2021, from That Data Tho...: https://thatdatatho.com/assumption-checking-lda-vs-qda-r-tutorial-2/