**Predicting Covid-19 Safety Regulation Compliance**

Jacqueline Ramos

Stevens Institute of Technology

EM 626 Applied AI & Machine Learning for Systems and Enterprises

Instructor: Dr. Carlo Lipizzi

TA: Denisse Martinez-Mejorado

**Abstract**

The COVIDiSTRESS Global Survey collected data on individuals’ responses to social science surveys in regards to stress caused by the COVID-19 pandemic and the individuals’ resulting behavioral responses. I used this dataset to create a classification model in order to predict compliance levels in regards to government regulated COVID-19 safety measures. The predictive model suffered from the imbalance of the dataset, with a vast majority of compliance level 4 frequencies, that oversampling and undersampling could not successfully adjust. However, the decision tree visualization did portray some correlation between the social science survey responses and an individual’s compliance level, as well as one’s country of residence and an individual’s compliance level. The survey depicting one’s concerns regarding covid and the survey depicting one’s trust in the media regarding covid both were found to be highly indicative of class prediction in the classification model decision tree. Initially I used a multiclass classification model to predict the compliance level as an integer between 1 and 6, as this was the ranking used in the survey. However, I found that using a binary classification, which predicts low compliance (compliance levels 1-3) and high compliance levels (compliance levels 4-6), yields the highest model accuracy, 0.83. I believe these stress level surveys can be indicative of compliance with safety measures, but further research should be done to better tune the model and understand its implications.

**Introduction**

Since the outbreak of COVID-19, there has understandably been a push towards research, studies, and data collection regarding the pandemic to better understand the virus and its various implications. With so many unknowns surrounding the pandemic, people are attempting to use pandemic data to identify patterns and trends in the hopes of helping people and industries rebuild. The utilization of A.I./Machine-Learning models could provide valuable tools and insights to this cause. I wanted to contribute to these COVID-19 efforts by trying to build a COVID-19 related predicative model, specifically in relation to the mental health and behavioral responses we have seen as a result of the pandemic.

## Background

Following the 2020 global outbreak of Covid-19, the world was sent into chaos, panic, and lockdown. In many countries, work, school, and public places were closed and people were urged to stay home unless necessary in attempt to slow the spread of the virus. Many people and businesses suffered economically. The mandate of wearing masks and social distancing in public was widely implemented. With most people thrust into a new way of living on top of dealing with the fear of a global pandemic, anxiety and stress were noticeably heightened. In an attempt to better understand the psychological and behavioral responses to the Covid-19 pandemic and one’s corresponding government safety efforts, the COVIDiSTRESS Global Survey was globally distributed over the months of April and May in 2020. This survey contained over 100 questions. It collected demographic information, as well as responses to various social science surveys. This information is detailed below.

### Demographic Information

The demographic information collected is as follows: age, gender, education level, mom’s education level, employment status, country of residence, expat status, state, marital status, number of dependents, riskgroup (for coronavirus) status, isolation situation, number of adults in same isolation place, and number of kids in same isolation place.

### Asian Disease Scenario Survey (AD)

This is a well-known social science survey that presents a hypothetical disease scenario and assesses an individual’s hypothetical reaction to various safety programs. Individuals choose their preferred safety measures.

### Perceived Stress Scale Survey (PSS)

This is a well-known social science survey that measures an individual’s stress. Individuals answer questions with a ranking on a scale from 1-5, 5 being the highest.

### SLON Survey (SLON3)

This survey measures an individual’s perceived loneliness. Individuals answer questions with a ranking on a scale of 1-5, 5 being the highest.

### OECD Trust Survey (OECD)

This survey measures an individual’s level of trust in other individuals and in institutions. Individuals answer questions with a ranking on a scale from 0-10, 10 being the highest.

### Corona Concerns Survey (Corona\_concerns)

This survey measures an individual’s level of concern over COVID-19. Individuals answer questions with a ranking on a scale from 1-6, 6 being the highest.

### Trust in Country’s Measures Question (Trust\_countrymeasure)

This question measures an individual’s trust level, on a scale of 0-10 and 10 being the highest, with their country’s COVID-19 safety measures.

### Compliance Survey (Compliance)

This survey measures an individual’s level of compliance with safety measures in their country. Individuals answer questions with a ranking on a scale of 1-6, 6 being the highest.

### BFF Survey (BFF)

This survey measures and individual’s perception of themselves, their characteristics, and how they handle stress. Individuals answer questions with a ranking on a scale of 1-6, 6 being the highest.

### Distress Survey (Expl\_Distress)

This survey measures an individual’s level of distress in regards to COVID-19. Individuals answer questions with a ranking on a scale of 1-6, 6 being the highest. If an individual feels the question does not apply to their situation they respond with ‘99’.

### Social Provisions Scale Survey (SPS)

This survey measures an individual’s available social provisions. Individuals answer questions with a ranking on a scale of 1-6, 6 being the highest.

### Coping Survey (Expl\_Coping)

This survey measures an individual’s level of coping with COVID-19 stress. Individuals answer questions with a ranking on a scale of 1-6, 6 being the highest.

### Media Survey (Expl\_media)

This survey measures an individual’s trust in the media. Individuals answer questions with a ranking on a scale of 1-6, 6 being the highest.

The responses to these surveys were cleaned and consolidated by data scientists into a csv dataset. I manipulated this dataset to create a predictive model for an individual’s compliance level. I chose to predict compliance levels because COVID-19 safety measures are a huge source of controversy, with some dutifully complying, some firmly refusing to comply, and some combination in the middle. Public figures urge and even beg people to follow the regulations. Those opposed have conducted protests in defiance of the regulations. There have even been viral videos of people refusing to wear masks on airplanes and being escorted off the flight by security. There are several different stances on and responses to the COVID-19 safety measures implemented by state and federal governments, so I thought it would be interesting to attempt to predict one’s likelihood of compliance based on demographics, attitudes towards COVID-19/ the country’s safety efforts, and an individual’s overall stress levels.

\*Additional information on the COVIDiSTRESS Global Survey research conducted can be found at links provided in the References section of this report.

## Methodology

### Cleaning the Dataset

I created my predictive program in Visual Studio Code using the python libraries sklearn, imbalanced learn, and pandas. I began by reading the COVIDiSTRESS Global Survey response dataset into a pandas dataframe. I then began cleaning the data as there was over 150 columns and 125,000 entries. I created a dataframe containing only the ‘Yes’ value in the answered\_all column to filter out the people who did not fully complete the survey questions. I then filtered out columns I felt were extraneous and not relevant to the compliance model like survey duration, recorded date, and the open text question responses.

### Computing Survey Averages

For the social science surveys, each question within each survey had a corresponding column where the ranking was recorded. These responses accounted for the majority of the columns in the dataset. In order to consolidate the survey responses, I calculated the mean response for each survey and stored this information in a new column. For example, for each entry, the 5 responses recorded for the Corona concern survey questions were averaged and the value added to the ‘Concerns\_avg’ column.

### Categorical Encoding

I then had to deal with the columns that had categorical values instead of numerical values. Sklearn’s decision tree classifier can only take in numerical values. In order to include the columns with categorical values, I had to encode them from categorical to numerical values. I used the library ‘category\_encoders’ to encode each categorical column. For example, the gender column has three values: ‘Male’, ‘Female’, and ‘Other/Prefer not to answer’. Category\_encoders assigns numerical values to each of these categories, like 1 for ‘Male’, 2 for ‘Female’, and 3 for ‘Other/Prefer not to answer’. Upon further research, I found that using category\_encoders to encode the categorical data is not sufficient for a decision tree classifier, as the classifier’s algorithm may interpret ‘Female’ as a value greater than ‘Male’ since it was assigned 2 and ‘Male’ was assigned 1. To combat this issue, I used another sklearn package ‘OneHotEncoder’ on the encoded categorical columns. OneHotEncoder creates a column for each of the different categorical values and assigns a Boolean binary value (0 if false and 1 if true) based on whether that entry contains that value or not. This method addresses the algorithmic issues of categorical encoding, but adds some computational expense as multiple columns are added in place of the original column. This isn’t too much of an issue for the gender column that only has three values and thus only adds three columns (and removes the original gender categorical column). However, the country column has over 100 values and so therefore adds over 100 columns when encoded with OneHotEncoder. I still chose to keep the encoded country columns despite the computational expense, as I was curious to see how one’s country affects compliance scoring. I encoded all the categorical columns then added them into a new dataframe (df\_enc) along with the rest of the numerical columns.

### Building the Classifier

Following categorical encoding, I was ready to build my classifier. I chose to use sklearn’s Decision Tree as my classifier. Data from the following columns were used as features:

|  |  |
| --- | --- |
| **Demographic Data** | **Survey Data** |
| Dem\_age | AD |
| Dem\_gender | PSS10\_avg |
| Dem\_edu | SLON3\_avg |
| Dem\_employment | SPS\_avg |
| Country | OECD\_avg |
| Dem\_expat | Concerns\_avg |
| Dem\_maritalstatus | Distress\_avg |
| Dem\_dependents | Coping\_avg |
| Dem\_riskgroup | Media\_avg |
| Dem\_isolation | Trust\_countrymeasure |
| Dem\_isolation\_adults |  |
| Dem\_isolation\_kids |  |

*Figure 1: Feature Table*

A full list of the features columns (with all the encoded columns) can be found under the Figures section in this report.

I used the compliance average column ‘Comp\_avg’ as my target variable. Thus, the model would be a multiclass classification model with the classes being 1, 2, 3, 4, 5, or 6. I then used the train\_test\_split method to split the dataset into training data and test data (about 1/3 of the data). I created the DecisionTreeClassifier instance and ran this data through it. I used graphviz to create a visualization of the decision tree. In order to analyze the model, I ran a classification report, checked the accuracy, and printed an error matrix.

### Tuning the Model

After creating the model, I played around with different methods to tune it and improve its accuracy. I created a column for compliance level buckets (‘Comp\_bucket’) with 1 representing a score of 1-3, 2 representing a score of 4, and 3 representing a score of 5-6. Additionally, I created a column for a binary compliance threshold value (‘Comp\_thresh’) with 1 representing a low compliance score of 1-3 and 2 representing a high compliance score of 4-6. I interchanged these columns as the target variable to see if the model’s accuracy increased. I also adjusted the decision tree classifier attribute values for class\_weight. Lastly, I used the oversampling and undersampling features in the library ‘Imbalanced Learn’ to address imbalance within the dataset. These efforts and their effect on the model are discussed in the following Results and Discussion sections of this report.

## Results

### Compliance Average as Target Variable (no weighting or over/undersampling)

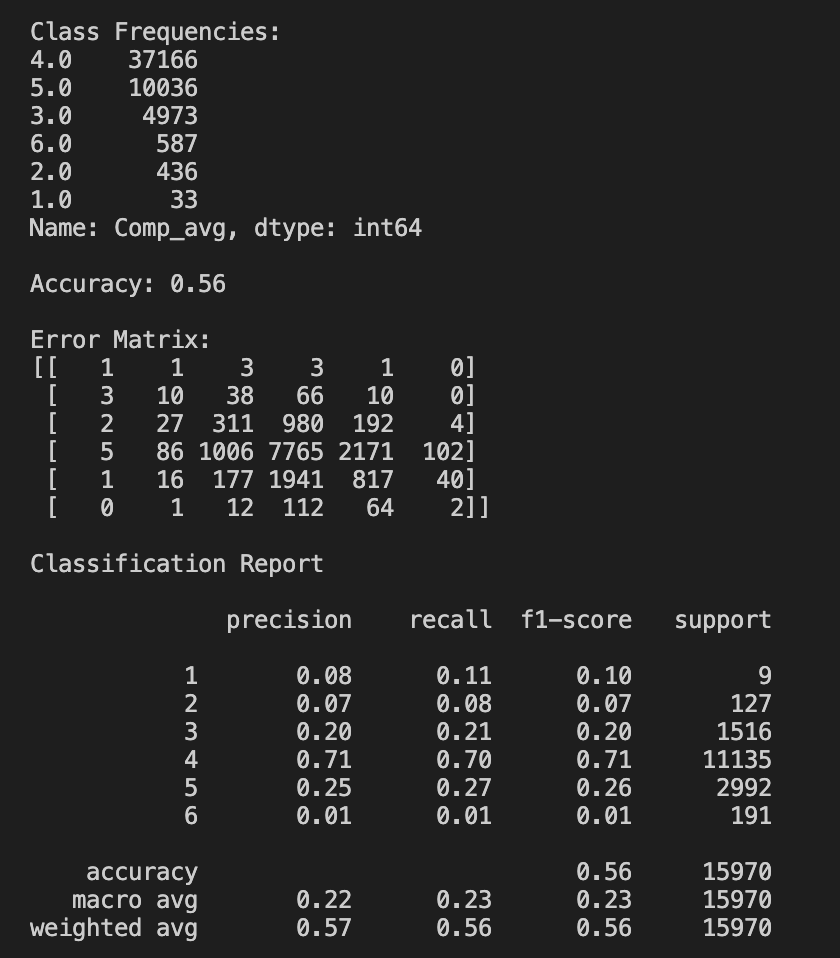
The figure below shows the analysis results for the classification model using compliance average as the target variable. Accuracy for the model is 0.56. Precision and recall are high for class 4 (0.71 and 0.70 respectively) but low for the rest, especially class 1, 2, and 6. This could be attributed to the imbalance of the dataset. The class frequencies show that there is a much higher number of occurrences of class 4 (37,166) than the rest of the classes. The error matrix demonstrates this, as well, with many instances of classification for class 4 and the majority of class 4 being classified correctly (7,765 correct classifications). However, this is not the case for the rest of the classes. For example, there are only seven instances of class 1 and only one is classified correctly.

Figure 3: Model Results with Compliance Average as Target Variable

### Compliance Average as Target Variable (with weighting and over/undersampling)

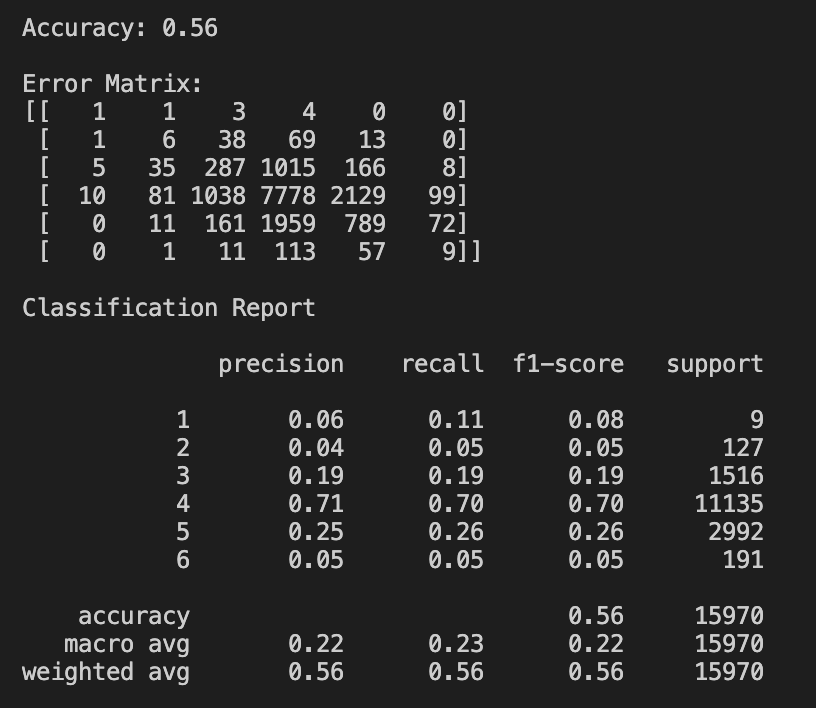
To address this imbalance, I set the decision tree classifier class\_weight attribute equal to “balanced”. This should adjust weights inversely proportional to class frequencies. However, upon running the model, no significant adjustments were made and, overall, most precision and recall scores reduced. This can be observed in the figure above. Accuracy remained 0.56 and only class 6 precision and recall slightly improved from 0.01 to 0.05.

Figure 4: Model Results with Compliance Average as Target Variable and Balanced Class Weight

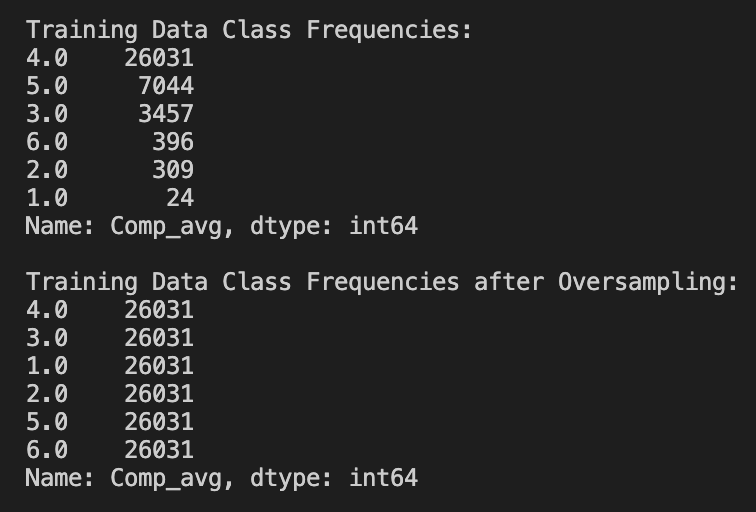
 I then attempted to use oversampling with Imbalanced Learn to duplicate minority class instances. I used the RandomOverSampler method with sampling\_strategy=”not majority” on the training data. This should duplicate the non-majority classes until there are equal occurrences of each class. The above figure shows the training data class frequencies before and after oversampling. After oversampling, all classes have the same frequency as majority class 4 (26,031).

Figure 5: Training Data Class Frequencies Before and After Oversampling

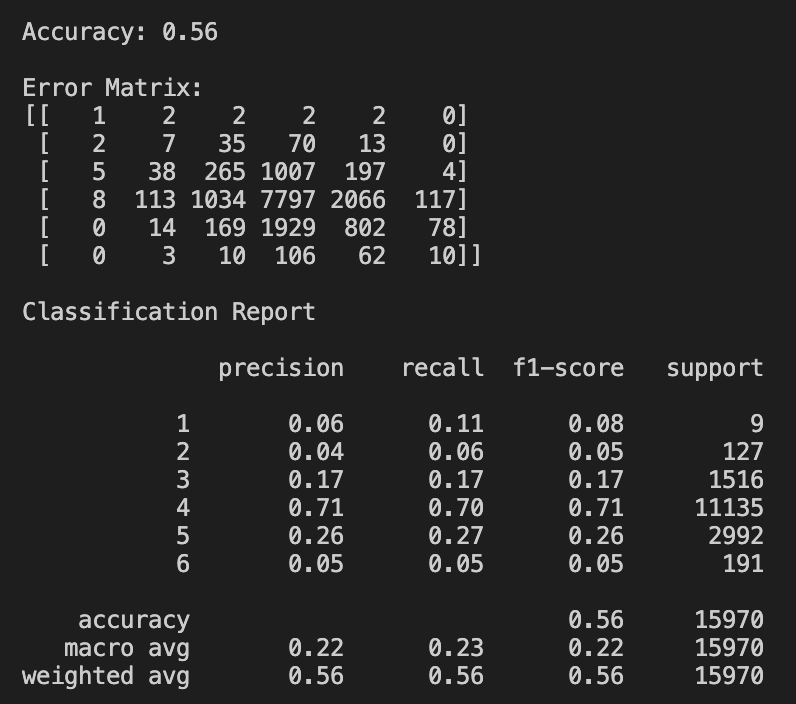
However, this oversampling did not significantly improve the model. Adjustments were minor as can be seen in the model analysis figure to the left. Accuracy is still 0.56 and precision and recall for each class is pretty much the same.

Figure 6: Model Results with Compliance Average as Target Variable and Oversampling

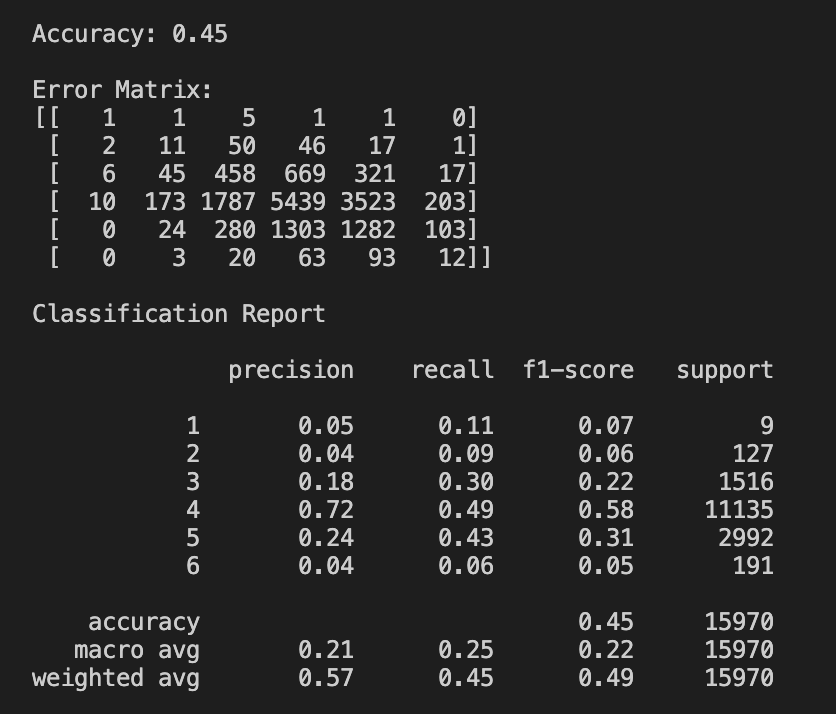
I then attempted using a combination of both oversampling and undersampling. I used the RandomUnderSampler method with sample\_strategy={4:10000}. This would bring the number of class 4 occurrences down to 10,000. I kept the same values as previously for the oversampling method. As can be seen in Figure 7, accuracy was reduced from

Figure 7: Model Results with Compliance Average as Target Variable and both Under/Oversampling

0.56 to 0.45, while precision and recall remained pretty much the same.

### Compliance Buckets as Target Variable

After playing around with different over/undersampling values but failing to significantly improve the model, I decided to alter the target variable. Instead of using the compliance score averages as the target variable, I decided to create three data buckets that grouped the compliance averages with 1 representing a compliance mean of 1-3, 2 representing a compliance mean of 4, and 3 representing a compliance mean of 5-6. I made these buckets in the hopes of more evenly distributing the data, as class 4 has the majority occurrences. The buckets only produce three classes instead of six.

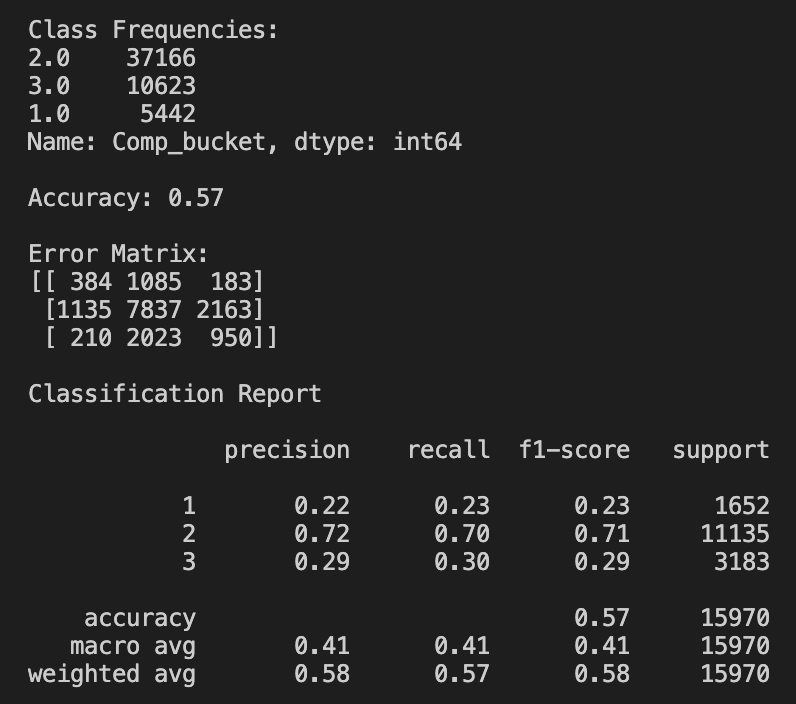
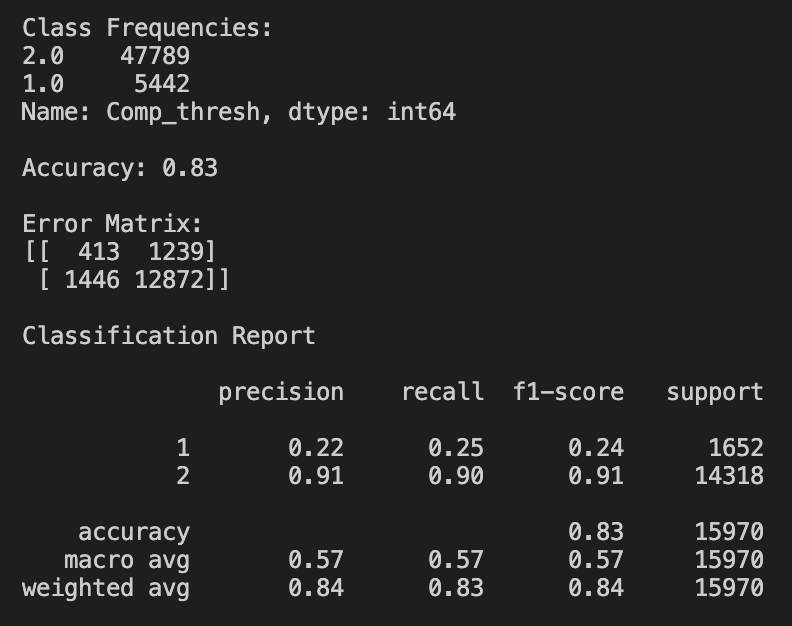
As can be seen in the figure to the left, the accuracy slightly increased from 0.56 to 0.57. With only three classes, the distribution of the data is better than previously. The overall precision and recall scores of the minority classes are improved (like 0.22 and 0.29) from the low scores of the previous model (0.04, 0.05, etc.). Class 2, which represents compliance average scores of 4, is still

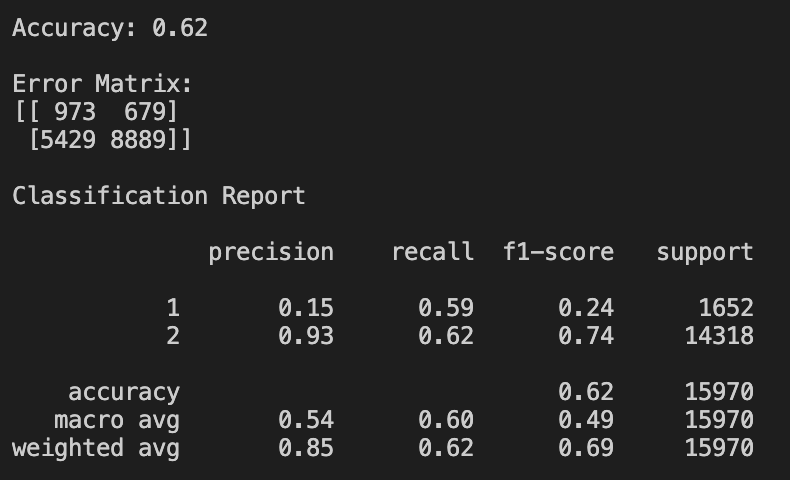
Figure 8: Model Results with Compliance Buckets as Target Variable

the majority class as can be seen in the error matrix distribution and the class frequencies (37,166 instances of class 2). I attempted to adjust the model with class\_weight values and over/undersampling, but, similar to the effects previously, they did not yield significant improvements.

**Compliance Threshold as Target Variable**

I also tried using binary classification instead of multiclass to see if it more evenly distributed the training data and raised the model accuracy significantly. I separated the compliance averages into 1, low compliance (compliance mean of 1-3), or 2, high compliance (compliance mean of 4-6). As can be seen in the figure to the left, overall accuracy of the model increased significantly to 0.83. Precision and recall are in the low 90s for class 2 and low 20s for class 1. There is still class imbalance with class 1 having 5,442 occurrences and class 2 having 47,789 occurrences.

Figure 9: Model Results with Compliance Threshold as Target Variable



I then attempted both over and undersampling to help with the imbalance. Overall accuracy was lowered to 0.62. Class 2 recall decreased to 0.62. Class 1 precision decreased to 0.15, but recall significantly increased to 0.59. Additionally, the error matrix identified

Figure 10: Model Results with Compliance Threshold as Target Variable and Over/Undersampling

the majority of instances of class 1 correctly and the majority of instances of class 2 correctly. A visualization of this error matrix and the error matrix of the model without over/undersampling can be seen in the figures below.

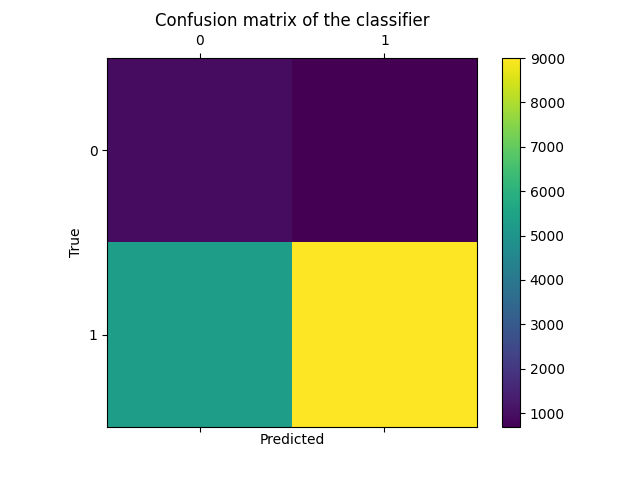


Figure 11: Confusion Matrix of Model with Over/Undersampling

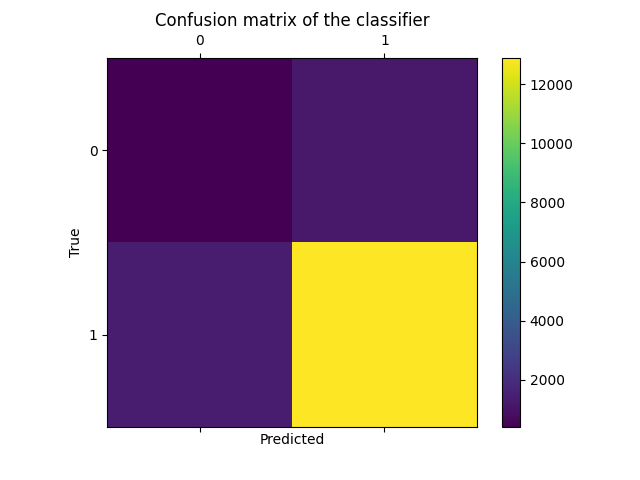


Figure 12: Confusion Matrix of Model without Over/Undersampling

In comparing the two figures, specifically in regards to class 1, Figure 11 shows a slightly lighter purple (meaning more occurrences), in the upper left corner, for correctly predicting class 1 and a darker purple (meaning less occurrences), in the upper right corner, for incorrectly predicting class 1. Figure 12, the error matrix without over/undersampling, shows a darker purple (meaning less occurrences), in the upper left corner, for correctly predicting class 1 and a slightly lighter purple (meaning more occurrences), in the upper right corner, for incorrectly predicting class 1.

Since using binary classification without over/undersampling yielded the highest accuracy of 0.83 (see Figure 9), I chose to visualize this model’s decision tree for further analysis. The full decision tree is quite large and time-consuming to produce, so I used max\_depth=4. A pdf of this figure will be attached to the report for clearer analysis.



Figure 13: Compliance Classifier Decision Tree

As can be seen above, the top node is the Country\_Japan column, which indicates whether the individual taking the survey is in Japan. My classification model found this to be the most indicative feature of class. Other highly indicative features include the Corona Concerns Survey column and the Media Survey column. These implications, as well as the implications of the rest of the results, are discussed further in the Discussion section.

## Discussion

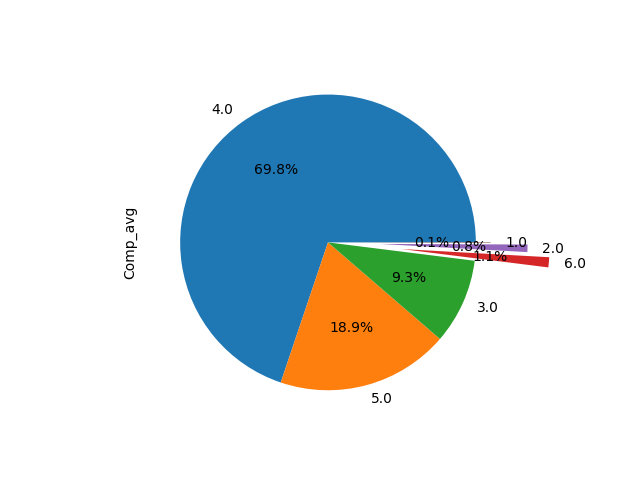
In analyzing the results, I found the biggest issue I was dealing with that was impacting my predictive model was the imbalance of the dataset. A vast majority of compliance averages are 4. The breakdown of the compliance average frequencies are in the figure below.

Figure 14: Compliance Average Frequencies Pie Chart

A compliance average of 4 makes up about 70% of the dataset entries. This imbalance seemed to skew the classifier model to overly predict class 4, when using Comp\_avg as the target variable. This model’s accuracy was 0.56 with the precision and recall of class 4 being fairly high, but the precision and recall of the other classes being alarmingly low. The error matrix, seen in Figure 3, also shows the majority of misclassifications of the other classes being class 4.

My attempts to rectify the data imbalance through class weighting and over/undersampling were futile. Balancing the class weights in the classifier adversely proportional to the class frequencies yielded no significant changes. Similarly, oversampling the minority classes to the same frequency as class 4 also yielded no significant changes. I tried various combinations of oversampling and undersampling (reducing the instances of class 4 and increasing the instances of the other classes), which surprisingly significantly reduced the model accuracy down to the 40s. However, a combination of both over/undersampling did increase the recall of both class 3 and class 5. Perhaps more time trying different over/undersampling methods could yield better results, but I was not able to achieve a useful combination within my time for this project.

I thought the model could still be improved, so I attempted to redistribute the data into groups or ‘buckets’ with class 1 being a compliance mean of 1-3, class 2 being a compliance mean of 4, and class 3 being a compliance mean of 5-6. I put class 4 in a bucket by itself because of its large number of occurrences. I used this bucket data as the target variable for my classifier. Class 2 (compliance mean of 4) was still vastly the majority of occurrences, so the accuracy only increased from 0.56 to 0.57. However, the precision and recall of the minority classes (class 1 and class 3) improved significantly from that of the minority classes without the use of data buckets. Looking at the error matrix, I see that class 2 (compliance mean of 4) is still being overly predicted for all classes. Although the overall accuracy only improved by .01, I believe this is a better predictive model than using compliance average without data buckets as the target variable because the data is a bit more evenly distributed. Before, compliance means of 1, 2, and 6 had severely low occurrences, but now they at least are grouped with other means, raising the overall class occurrences. Similar to the previous model, I was also unable to achieve improvement via over/undersampling or class weighting.

I was able to greatly improve model accuracy by further grouping the compliance average data and using binary classification. Class 1 represented low compliance with compliance means of 1-3 and class 2 represented high compliance with compliance means of 4-6. This data was still imbalanced with 47,689 instances of class 2 and 5,442 instances of class 1. However, the accuracy significantly improved to 0.83. The precision and recall of class 2 improved to the 90s and the precision and recall of class 1 was in the 20s. Looking at the error matrix, I still saw a majority of the minority class (class 1) being incorrectly classified as class 2. This must still be a result of class 2 containing the compliance mean of 4 and its large occurrences. Once more, I tried over/undersampling to balance the dataset. This resulted in an accuracy decrease to 0.62. However, the error matrix changed from previous patterns. This model’s error matrix had the majority of instances for each class being classified correctly. I achieved this by undersampling class 2 training data from 33471 to 4000, then oversampling class 1 training data from 3790 to match the majority class instances (4000). I found this to be quite interesting. I wonder if the previous binary model with no over/undersampling had such a high accuracy of 0.83 because there was less of a chance of the model predicting wrong with only two classes. This model seems to still be overly predicting class 2, which contains compliance mean of 4, but has a greater chance of getting the prediction right since this time there are only two data groups, instead of three or six. Maybe then, the better model is the binary one with over/undersampling, since the data is more evenly distributed, even though the accuracy is lower. Perhaps more research on data imbalance and the benefits of over/undersampling could be useful.

I also find the decision tree visualization to be quite interesting. Country\_Japan was the top node in the tree and thus an important correlation to compliance prediction. Prior to creating the predictive model for this project, I thought one’s country of residence would be highly indicative of compliance level. Some countries have strict COVID-19 regulations, which many may not comply with. Other countries have fewer safety measures that may be “easier” to follow. However, I find it interesting that the country, Japan, was the top node as opposed to the many other countries. Upon some online research, I found the COVID-19 regulations in Japan to be on the more strict side, especially at the time of survey distribution. I think it would be interesting to conduct a data exploration in the future of just the Japanese entries in this dataset and compare it to research on Japan’s societal climate in regards to COVID-19. The other nodes towards the top of the decision tree were the survey averages. This shows they were also considered highly indicative of classifying compliance levels. The two highest were the Corona Concerns Survey and the Media Survey. This makes sense because if one is averaging high levels of concern in regards to the coronavirus, I would think this would be indicative of high compliance with the COVID-19 safety measures. If one is referring to the news and government media for guidance with COVID-19, resulting in a high media survey average, I would think they are complying with what the government is advising. I also thought it was interesting that Dem\_employment\_retired was high up in the decision tree. This makes sense to me because COVID-19 can be more severe and result even in death for the elderly population. Often, those who are retired are older and therefore more at risk. I would think they would have high compliance ratings since their risk is higher.

**Conclusion**

Overall, I think my predicative model provided interesting results and insight into COVID-19 trends that could prove helpful in understanding the controversy and chaos that surrounds COVID-19 safety regulation compliance. I believe there is a correlation between social science and COVID-19 compliance levels. However, I think the predicative model could still be improved to better portray this. The imbalance of the dataset hurt the model’s accuracy. I think further tuning of the model and research into under/oversampling techniques to combat the imbalance should be conducted. I also think isolating the dataset by country and running it through the predicative model could be useful for more insight. Future research on Japan in regards to COVID-19’s societal impact and the government’s regulations could be helpful in better understanding its place at the top of the decision tree.

References

*Covid-19 information*. U.S. Embassy & Consulates in Japan. (2021, December 8). Retrieved December 12, 2021, from https://jp.usembassy.gov/covid-19-information/.

Hundley, D. (2020, September 7). *Data science tip #005: Two correct ways to perform one-hot encoding!* Medium. Retrieved December 12, 2021, from https://towardsdatascience.com/data-science-tip-005-two-correct-ways-to-perform-one-hot-encoding-7c21e1db7aa2.

*Over-sampling methods¶*. Over-sampling methods - Version 0.8.1. (n.d.). Retrieved December 12, 2021, from https://imbalanced-learn.org/stable/references/over\_sampling.html#basic-over-sampling.

*Sklearn.tree.decisiontreeclassifier*. scikit. (n.d.). Retrieved December 12, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

Yamada, Y., Ćepulić, DB., Coll-Martín, T. *et al.* COVIDiSTRESS Global Survey dataset on psychological and behavioural consequences of the COVID-19 outbreak. *Sci Data* **8,**3 (2021). https://doi.org/10.1038/s41597-020-00784-9

\*Link to survey research article: <https://www.nature.com/articles/s41597-020-00784-9#Tab9>

**Additional Figures (not included in this report)**

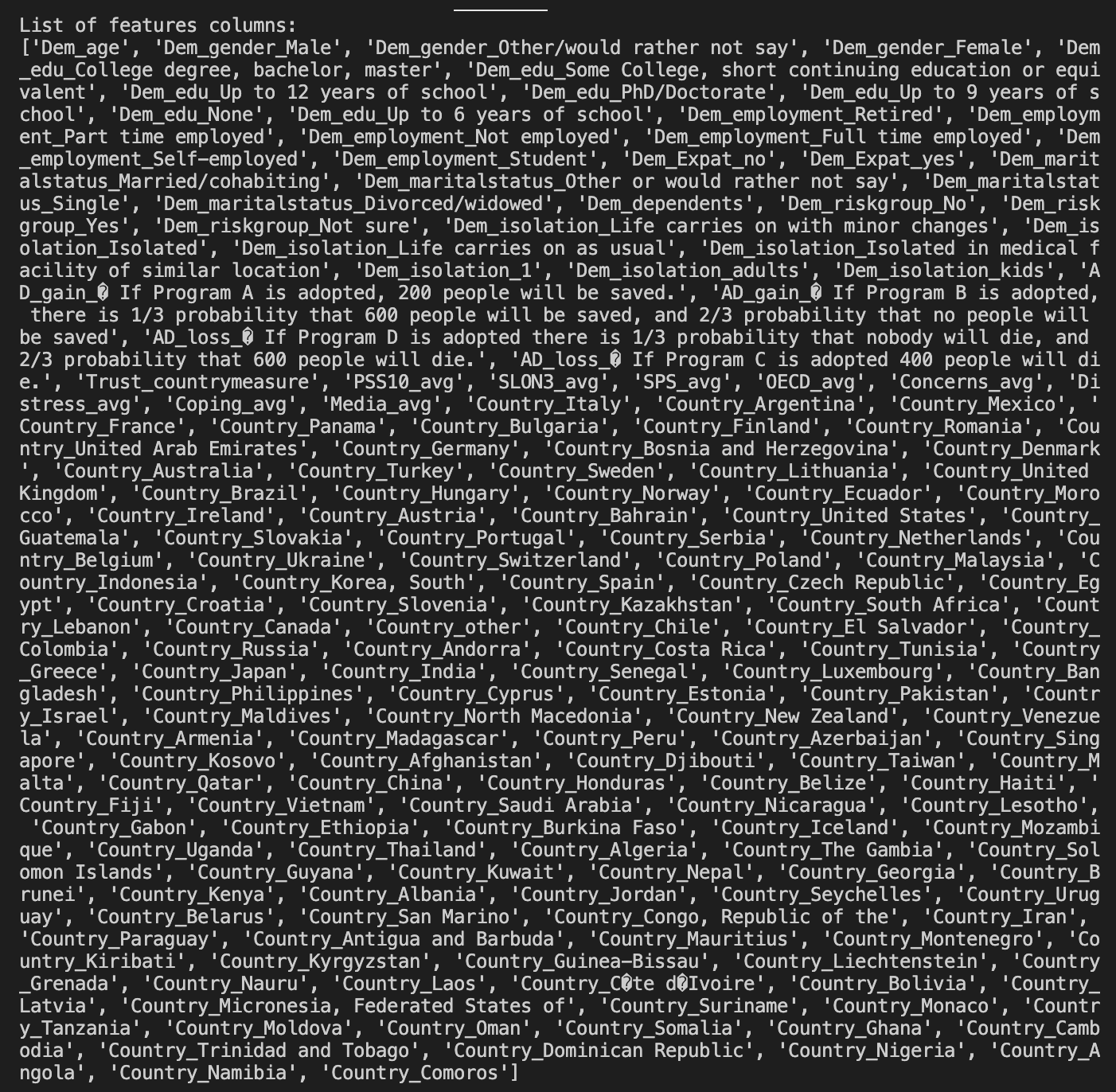


Figure 2: Full List of Features Columns