Responsible Data Science —— Final Report

Yichen Guo, Yue Wu

1. Background

The "ADS" we chose finds the correlations between race/ethnicity, gender, poverty, severe health conditions and Covid 19 risk index, morbidity and mortality.

The ADS has two phases, in the first phase:

The user **evaluates morbidity (Cases as indicator) and mortality (Deaths as indicator)** using a series of regression models (Ordinary Least Squares), within which each regression uses a different combination of features. The regression models were compared based on the statistical significance of features, AIC (Akaike information criterion) and R-square. The stated goal of the first phase is to "conduct explanatory analysis and test hypotheses that no relationship exists between features using the OLS regression."

In the second phase (log-transformed pre-processed data from phase one was used):

The user fitted Random Forest regression models on various training datasets with different features selected (all features, most important features, dropping tangling features) to **predict morbidity and mortality.** The stated goal of this phase is to "test hypotheses that features with highest importance are **unable** to predict Covid 19 morbidity and mortality using machine learning (Random Forest)."

After figuring out the Kaggle project is not an actual ADS, it's more like fitting data in to a given model (the Random Forest Model and OLS regression), so we discussed and decided to implement our own simple ADS (we will also use a Random Forest Classifier).

2. Input and output

2.1 Data used:

The raw data used are: *US Coronavirus Cases, US Coronavirus Deaths* (from 22/01/2020 to 31/07/2020), *State/County Poverty Universe Data, All ages*, and *Annual County Resident Population Estimate by Age, Sex, Race and Hispanic Origin* (US Census 2019) and *Severe COVID-19 Health Risk Index by U.S County*. The data used for the ADS was a consolidation of the above dataset. It uses COVID 19 risk, cases and deaths data as foundations and then adds other features based on geographical information.

The consolidated data was sourced on race/ethnicity, gender, poverty and severe health conditions and Covid 19 morbidity and mortality at the U.S county level. Then the final data used by the ADS was " cleaned and pre-process (log-transformed) according to unique identifiers." However, there's no detailed record of how the data were joined, cleaned and how the identifiers were selected.

The data provided on Kaggle were already cleaned and log-transformed with non missing data observed. The value of columns named Poverty, Population, Cases, Deaths and columns that represent demographic groups (e.i.: W_male, B_female) are all log-transformed population count data. For example, input of the Cases columns were calculated by *log(actual Covid 19 cases)*.

The table shows the data description of all columns (exclude self-explanatory column names: County and State):

Column Name	Description
FIPS	5-digit Federal Information Processing Standard Publication (FIPS) code for U.S. Counties (composite of stateFIPS and CountyFIPS)
stateFIPS	2-digit Federal Information Processing Standard Publication (FIPS) code for U.S. States
countyFIPS_2d	3-digit Federal Information Processing Standard Publication (FIPS) code for U.S. Counties
Cases	Cumulative Covid 19 cases in County (08/31/2020) (log-transformed)
Deaths	Cumulative Covid 19 deaths in County (08/31/2020) (log-transformed)
Poverty	Individuals (all ages) in the county classified as living in poverty (2019) (log-transformed)
Population	Total number of residents in the county (2019) (log-transformed)
W_Male / W_Female	Total number of residents in the county identified as White Male / White Female or combination with other (2019) (log-transformed)
B_Male / B_Female	Total number of residents in the county identified as Black Male / Black Female or combination with other (2019) (log-transformed)
H_Male / H_Female	Total number of residents in the county identified as Hispanic Male/Female or combination with other (2019) (log-transformed)
I_Male / I_Female	Total number of residents in the county identified as American Indian and Alaska Native Male / Female or combination with other (2019) (log-transformed)
A_Male / A_Female	Total number of residents in the county identified as Asian Male / Asian Female or combination with other (2019) (log-transformed)
NH_Male / NH_Female	Total number of residents in the county identified as Native Hawaiian and Other Pacific Islander Male/ Female or combination with other (2019) (log-transformed)
Risk_Index	Raw Risk Index score 0-100 (log-transformed)

Risk_Cat	Risk Category Name (Below Average, Above Average, HIgh, Low, Very High, Very Low)

We used some of the output from the exploratory analysis generated by Kaggle bot (we also add some codes for further observations). The picture below on the left hand side shows the column names and types. The picture on the right hand side shows the correlation heatmap (only columns with numeric attributes were shown). Figure 2 shows an overview of the correlations between all features that the ADS used during the analysis. By viewing the overall correlation, we can notice our "columns of interest", as well as how the user chooses specific columns to analyze. For example,we notice negative correlation between risk _index and cases and other features, while strong positive correlation between cases and poverty and black male/ females.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3142 entries, 0 to 3141
Data columns (total 23 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   FIPS           3142 non-null   int64
 1   stateFIPS      3142 non-null   int64
 2   countyFIPS_2d  3142 non-null   int64
 3   County         3142 non-null   object
 4   State          3142 non-null   object
 5   Cases          3142 non-null   float64
 6   Deaths         3142 non-null   float64
 7   Poverty        3142 non-null   float64
 8   Population     3142 non-null   float64
 9   W_Male         3142 non-null   float64
 10  W_Female       3142 non-null   float64
 11  B_Male         3142 non-null   float64
 12  B_Female       3142 non-null   float64
 13  H_Male         3142 non-null   float64
 14  H_Female       3142 non-null   float64
 15  I_Male         3142 non-null   float64
 16  I_Female       3142 non-null   float64
 17  A_Male         3142 non-null   float64
 18  A_Female       3142 non-null   float64
 19  NH_Male        3142 non-null   float64
 20  NH_Female      3142 non-null   float64
 21  Risk_Index     3142 non-null   float64
 22  Risk_Cat       3142 non-null   object
dtypes: float64(17), int64(3), object(3)
memory usage: 564.7+ KB
```

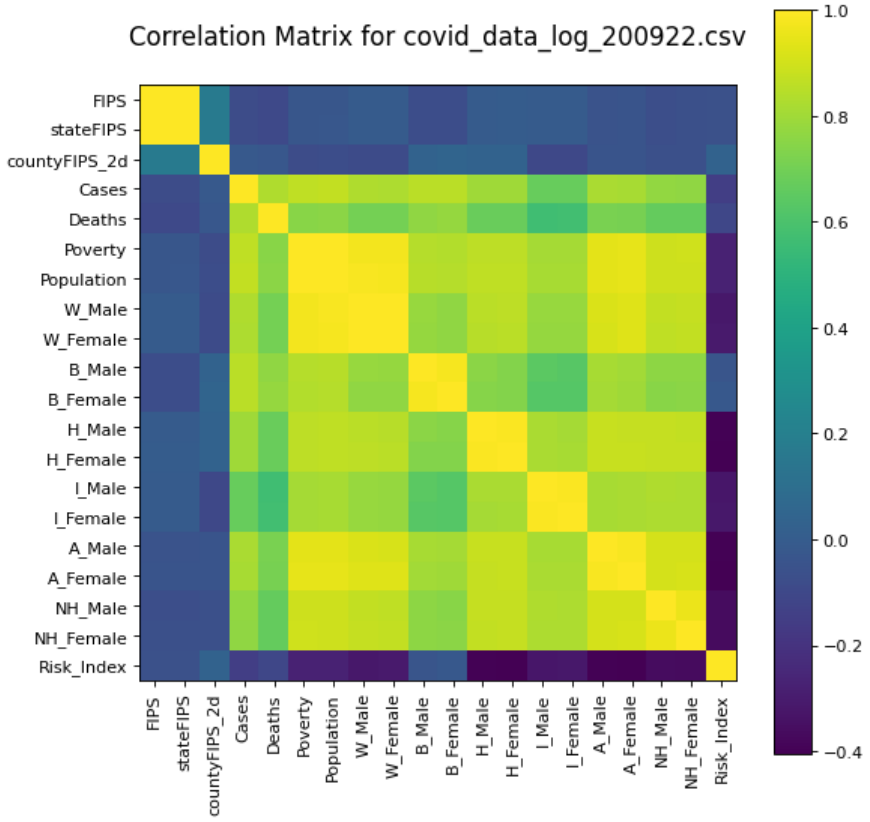

Figure 1: Summary and correlation matrix of "covid_data_log_2009022"

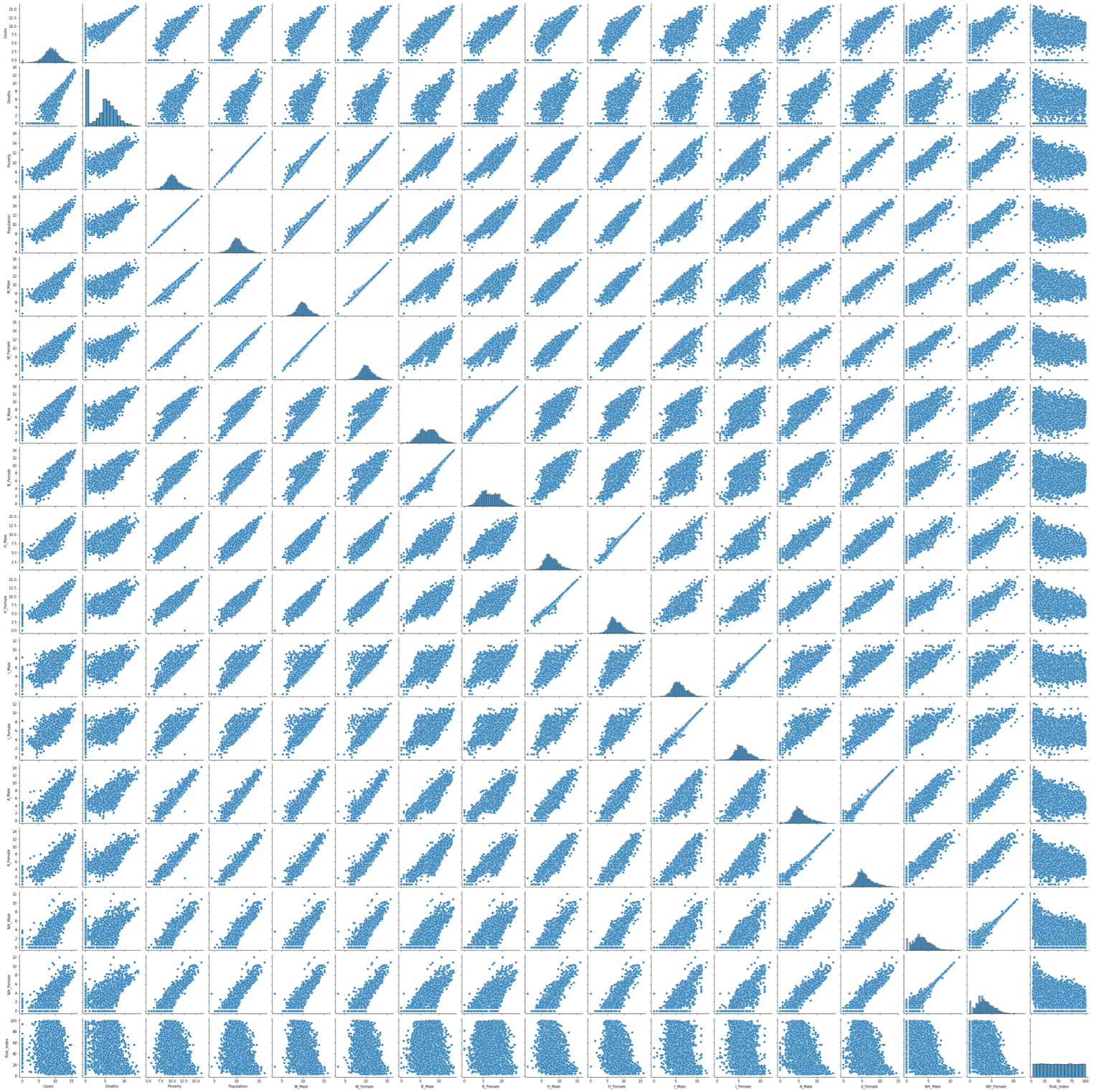

Figure 2: Correlation table

2.1.1 Distribution of demographic data (our approach to our ADS)

Since the data were already cleaned and pre-processed, the demographic distributions of the demographics data are close to normal distributions (see figures below). The distribution of male and female also follows a similar pattern among all demographic groups.

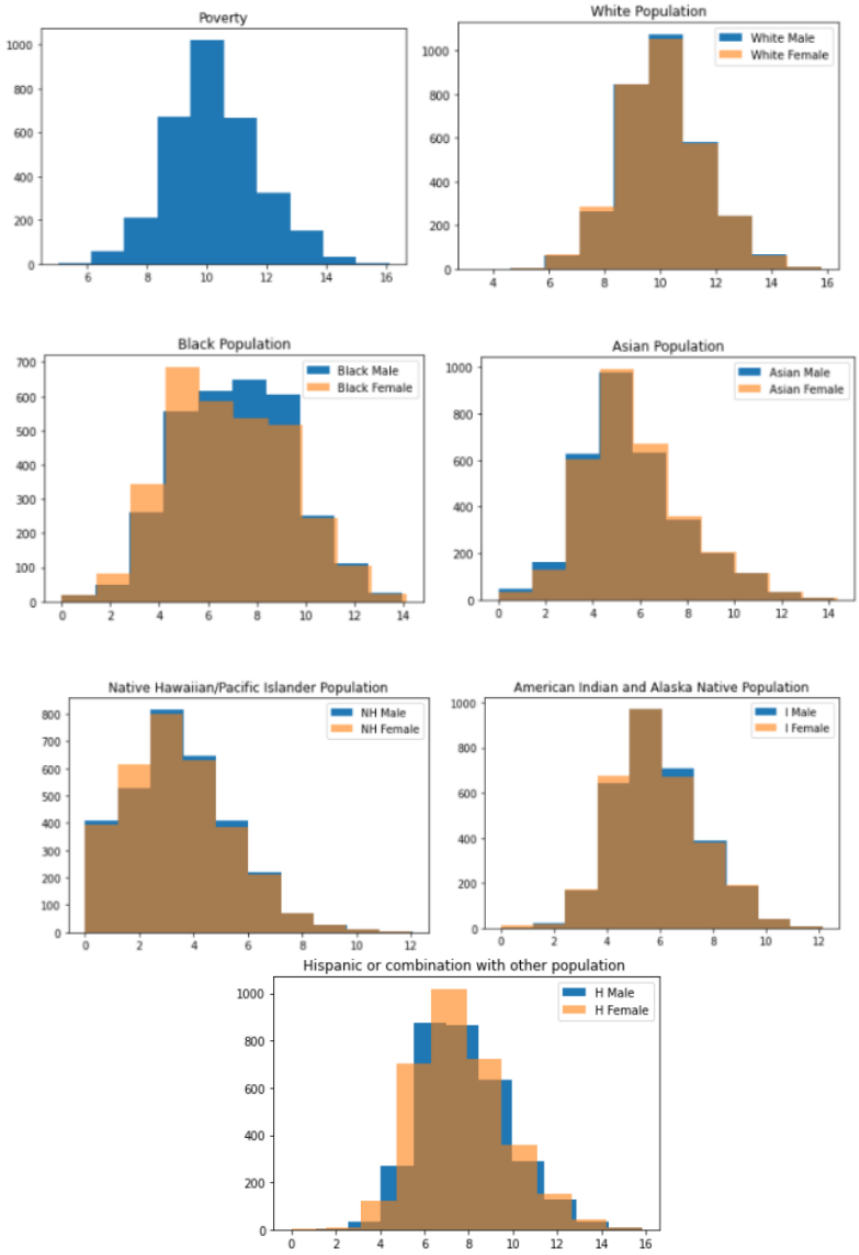

Figure 3: Distribution of population of poverty and demographic groups

From figure 4, we can see that white accounts for the largest portion of total population, followed by the hispanic and black population. Only a minor difference was observed between American Indian/Alaska Native and Asian, while Native Hawwaiian/Pacific Islander accounts for the smallest portion of the whole population. From the distribution of population among different ethnicities, we can conclude that the data used was reasonable and inclusive that was not in favor of any racial group by giving an unreasonably large/small portion in the whole population.

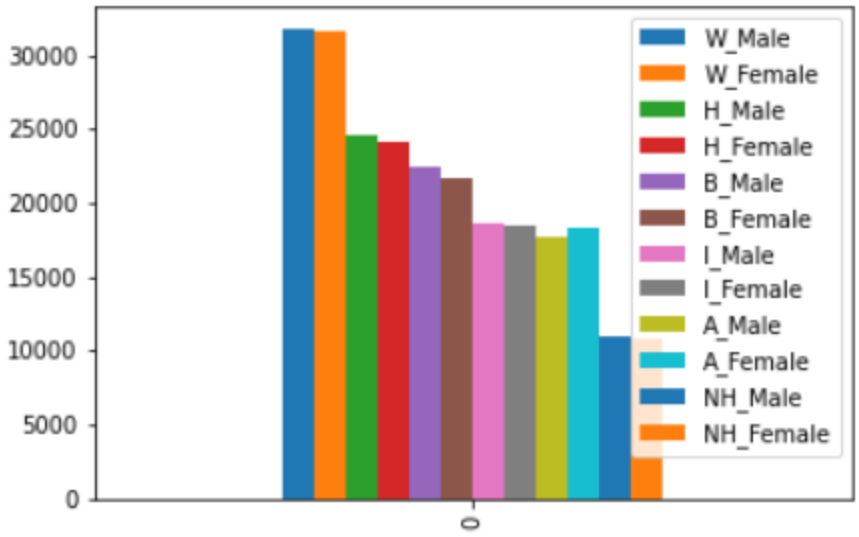

Figure 4: Population count of all racial groups

2.1.2 Distribution under specific risk category:

Risk category here means the risk category of a county, not a specific racial group. So the graph represents the total population count of each racial-gender group of all counties that is labeled as a specific risk category. In other words, the graph indicates how much each racial-gender group contributes to a specific risk category in general. We looked at the distribution of the racial groups under each risk category, using white population as a reference.

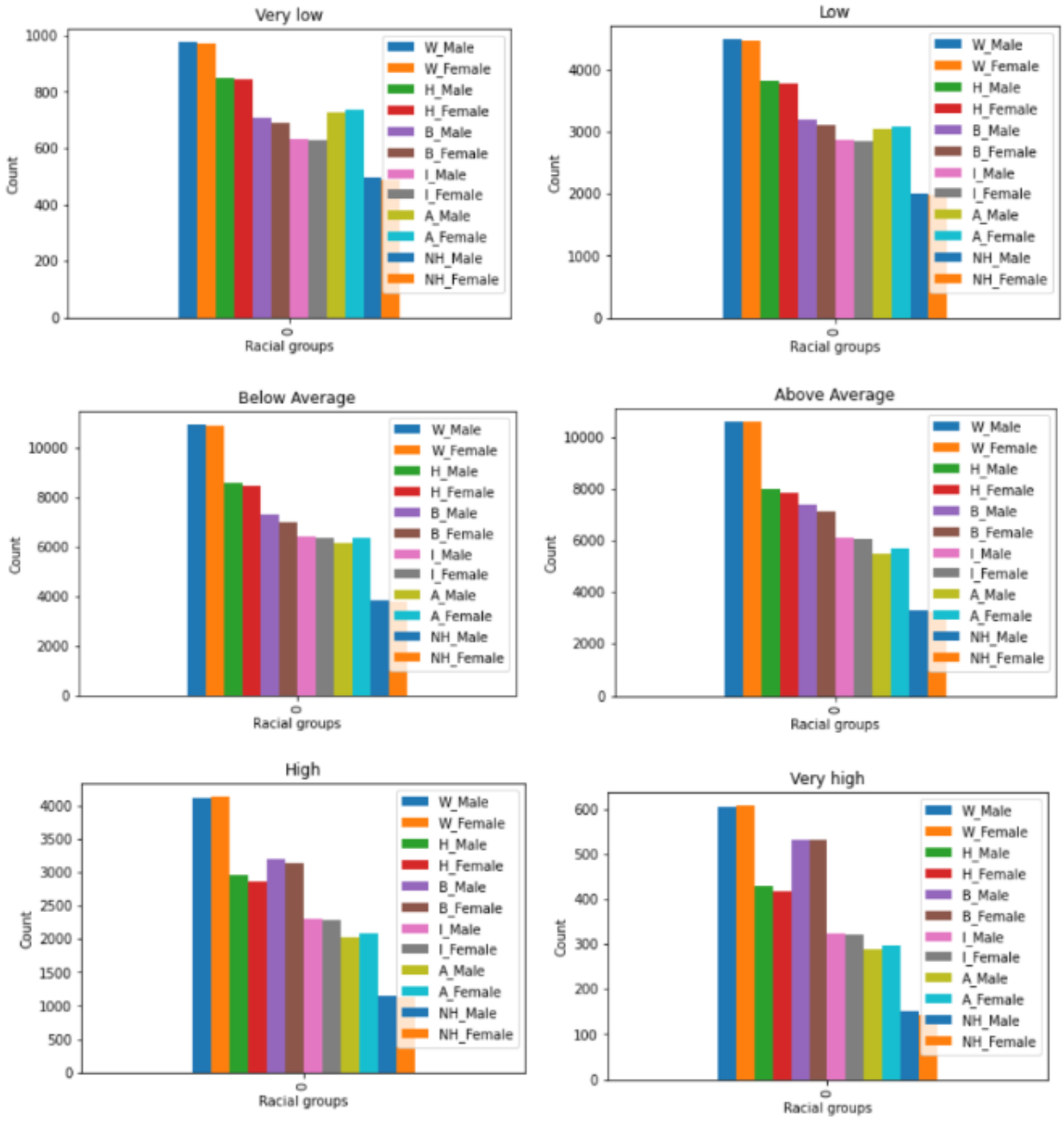

Figure 5: Population distribution of risk categories

While the main trend remains similar with the overall population distribution. There's a greater number of Asian populations when the counties' risk category is "very low" and slightly higher numbers of Asian populations when the risk category is "low". For the counties' risk

level of "high", we can see a moderate increase in black population. When it comes to the "very high" category, the increase in black population is more obvious compared to the original distribution. For the "very high" category, we can see that there's a decrease in distribution of American Indian/Alaska Native and Asian population.

Since the difference between populations of racial groups already existed, so looking at the portion of each racial group that contributes to a specific risk category is clearer and more interpretable/ Figure 6 shows the portion of a specific racial group under a certain risk category.

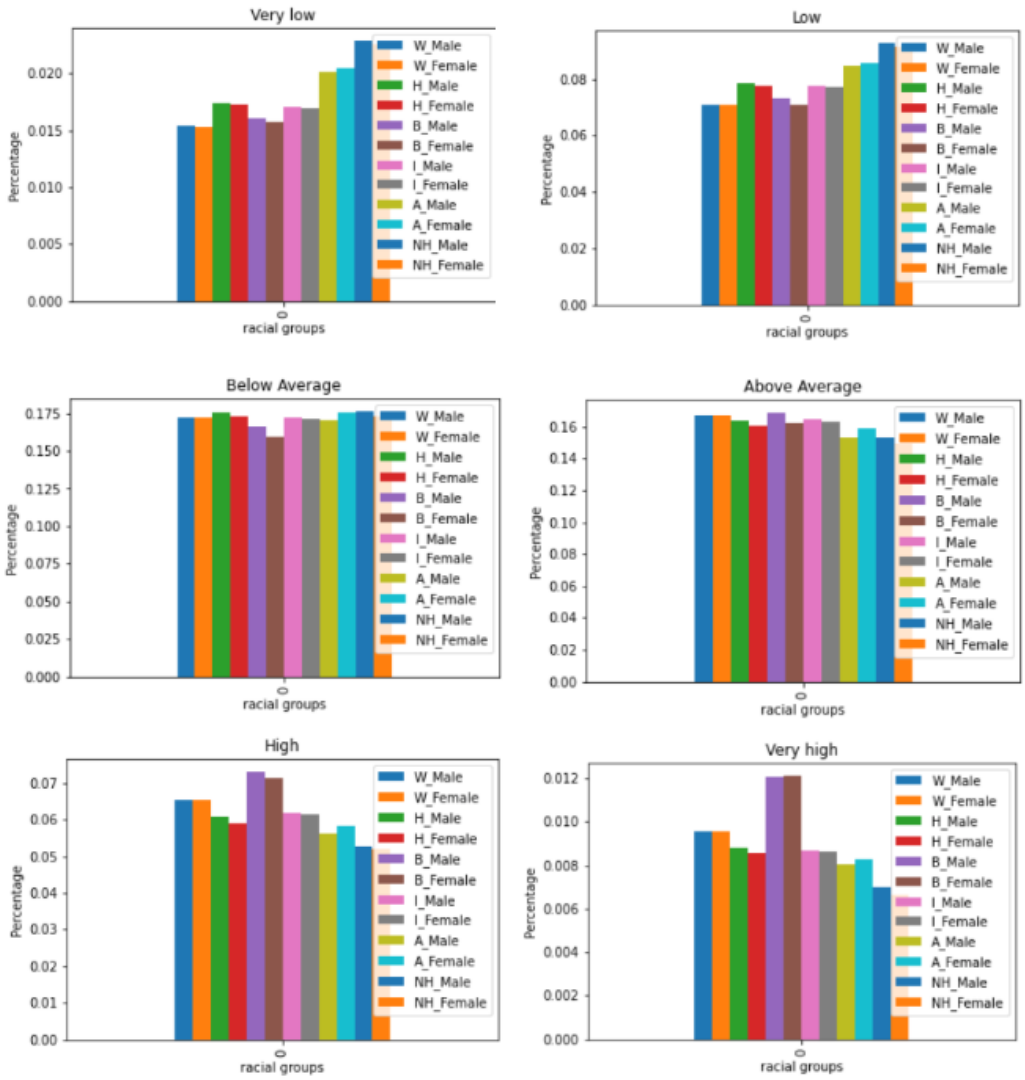

Figure 6: Portion of racial groups under given risk category

Ideally, if the risk category prediction is independent of racial groups, the racial group population makeup for each risk category should be similar. From figure 6, we didn't observe absolute discrepancy between male and female population regarding risk categories. We can see that a higher portion of Hispanic, American Indian/Alaska Native and Asian were observed under the risk categories of "very low" and "low". The "below average" and "above average" categories have almost ideal distribution. The greatest difference is observed with a high portion of black population standing out in the "high" and "very high" categories. We may infer from this observation is that when a county has more black population, it is more prone to be labeled at a higher risk category. By contrast, if a county has more American Indian/Alaska Native and Asian, it's more likely to be labeled at a lower risk category. This observation also fits the result from the Kaggle project that black population has more significance in determining both morbidity and mortality of a county.

Based on the observations and the fact that we will be generating our own simple ADS, we found it's interesting to look at the relationship between the population distribution of a county and the risk category of that county. We'll be focusing on whether the given data is biased towards black population leading to a high risk index. We use Case, Deaths, Populations and 12 demographic populations in the above data description form as inputs. These selected features will be used to predict the risk category (high/low).

2.2 The output

2.2.1 Kaggle output

The very early phase of the Kaggle user did some work in finding the relationship between COVID cases and the risk index (since a negative correlation was found in the overall correlation graph). The output is a numerical range of the coefficient of risk index with 95% confidence interval in predicting cases. The risk score can be interpreted as a score that reflects the county's risk during covid, with a higher score indicates a greater chance of higher morbidity or mortality.

The later phase of the user's project addresses the prediction of morbidity and mortality. The output of the system is predicting morbidity (cases) and morality (deaths) by finding the best model with the significant features, which returns a log-transformed value of the cases and deaths count. The user uses cases to represent morbidity and deaths to indicate mortality, under which more cases represent higher morbidity and more deaths indicate higher mortality.

2.2.2 Our output

We will be using the given processed data and use the risk category as the target outcome. For a simpler and clearer explanation, we divide the risk category into low and high risk using the median of the risk index. The risk category shows the COVID risk level of a county given its demographics makeup. The six categories in the original data are: very low, low, below average, above average, high and very high. Our ADS will have only two categories: low (0) and high (1).

3. Implementation and validation:

3.1 Implementation and validation of Kaggle project

The code comes from Kaggle, some data cleaning and processing was already done, including dropping unrelated columns, group individuals by counties, normalizing the statistics. The implementation of the system takes in numerous data from different aspects, mainly focuses on finding models that best predict morbidity and mortality using different features. The Kaggle user used the OLS regression (Ordinary Least Square) regression in training the data and found the correlation between cases and risk index. The user used a Random Forest Regressor in phase 2 with the feature importance method (out-of-bag = True) to find the best model in predicting morbidity and mortality.

In phase 1, the user used AIC (Akaike information criterion) and R-square to evaluate the models. R-square measures how close true data are to the predicted data. It's calculated by the percentage of the response variable variation (explained variation/ total variation). So the higher the R-square, the better the model fits the actual data. AIC is an estimator of prediction error and it's used to find the best model in a given set. It's calculated by AIC = 2k - 2ln(L_hat), which deals with both the risk of overfitting and underfitting. For each model, k is the number of estimated parameters, and L_hat is the maximum value of the likelihood function. So a lower AIC value indicates better fit (with high log-likelihood value and reasonable parameters used). In phase 2, the user used mean absolute errors in evaluating models for predicting morbidity and mortality, which is calculated by the sum of the absolute difference between predictions and true value, divided by the number of data points. So the smaller the MAE is, the better the model.

3.2 Our implementation and validation (since we are building our own model):

We used random forest classifiers to build our risk-classification tool, which gives us an accurate and stable prediction. We used SHAP to explain the model. For random forest, we are looking at both AUC and accuracy, for SHAP, we will be looking for feature importance.

Random Forest Classifier builds multiple decision trees and merges them together to get a more accurate and stable prediction, and we make use of it to predict the risk category with given features, such as cases, deaths, poverty, population and a bunch of demographic statistics. More specifically, we used GridSearchCV with given parameters settings and the parameters are optimized by cross-validated grid-search over a parameter grid. We used AUC for the scoring, which means the GridSearchCV will use AUC as an evaluation metric when finding the best model.

For validation, we focus on optimizing the AUC (Area Under ROC), since the higher AUC, the better the model is at distinguishing between high risk and low risk. That is, a better model would have a good measure of separability. The reason we chose AUC is that the model is predicting the risk category of a given county, more likely it will be suggestive to stakeholders like government decision makers, health care departments that will make COVID remediation decisions based on the risk level of a given county. The value we set as a threshold of our model's AUC is at least 0.8. While when the value of AUC is close to 0.5 indicates that the model is not doing well at differentiating the two classes.

The other metrics we also looked at but may not be prioritized when choosing a model are accuracy, precision (average precision) and recall (TPR). For accuracy is calculated by (TP+TN)/(TP+FP+FN+TN), which measures the percentage of correct prediction, we try to keep

accuracy close to 0.8. Precision is calculated by TP/(TP+FP), which shows the ratio of the corrected positive predictions among all positive predictions. In our case, precision shows the portion of the given data our model predicted to have high risk were indeed high risk. Recall is also known as the true positive rate, which represents the ability of our model to identify all positive instances, and is calculated by TP/(TP+FN).

SHAP is a game theory approach to explain the output of any machine learning model, with its values interpreting the impact of a particular value for a given feature on the prediction. In our case, we use it to explain how each demographic group impacts our target value, the risk category. We also used LIME as a backup explainer since LIME more clearly summarizes the weights of each feature and its chosen class.

For fairness among counties, we compared our model performance (accuracy, FPR, FNR and TPR (recall)) of counties with different demographic makeup. We looked at whether our model performs the same for counties with more populations of a specific gender, race and poverty. For example, a county with more male is defined as a county's Male to Female ratio is greater than the median Male to Female ratio of the overall populations. The ratio for a racial group is calculated by the population of that racial group divided by the sum of all racial groups' population. We also used a median as a threshold in deciding whether a county has more population of a given racial group. Median was also used to decide whether a county has more people in poverty. Since it's COVID19 related prediction, so FNR is really important since wrongly placing a high-risk county as low-risk may worsen the situation of that county since no proper remediations may be taken under the wrong suggestions. It's also important to look at FPR, since wrongly labeling a low-risk county as high-risk can lead to over-contributing resources that could be used to help counties under high-risk.

4. Outcomes

4.1 Outcomes of Kaggle project

The model that generated the final observation of the Kaggle project only used the female population data when training the model, so we can't tell how the model performs for the male population. However, there was a much higher importance of balck female in predicting morbidity and mortality. The outcome of the Kaggle model found that the model evaluated cases to risk index performed best in terms of AIC and R-squared. While case was not used as a feature, there's an increase in MAE value from 1.36 to 1.71 in predicting mortality. However, the MAE remained the same in predicting morbidity with/without case as a feature. The Kaggle project used the out of bag feature in the random forest model and found that the balck female is the most important feature in predicting morbidity.

4.2 Our outcomes

We learned a lot about random forest classifiers and SHAP as tools to help us analyze data and predict values. Despite encountering obstacles while implementing these tools, we manage to conquer those problems and to navigate to a satisfying result.

Nutritional Label:

4.2.1 Accuracy, AUC, Precision, Recall

The model's AUC was around 0.76 before tuning and it increased to 0.84 after tuning. For accuracy and other model performance metrics, they remained approximately the same before tuning.

The final best model we got is a tuned random forest model generated using the GrivdSearchCV function, which has an accuracy of 0.75, an AUC score of 0.841, a precision of 0.786 and a recall of 0.743.

4.2.2 SHAP and LIME

The SHAP explainer did not show significant results and no obvious pattern of feature importances in predicting risk category was observed. The points worth noting are that we did not see a heavy weight putting on Cases, Deaths and Population, and discrepancies were observed about the weights of racial groups in predicting risk categories. More specially, figure 7 shows 5 correct predictions explained by SHAP, for the first case, A_male was put at a heavy weight towards high-risk class, while B_Female was put at a heavy weight towards low-risk class. However, in the second example, A_male had a heavy weight in favoring the low-risk class while B_Female was in favor for the high-risk class. Below the explanations is a summary plot which also didn't show any trend of any feature favored one class over the other.

We also tried LIME which we expected to discover some trends of the input features. By comparing the explanations in figure 8, we speculate that the feature's favored class is related to the value of that feature. For example, in the first example with black female population (log-transformed) of less than 5, it's favored towards the low-risk class, while in the second example, when black female population is greater than 8.78, it's favored towards the high-risk class. The same pattern is also observed with Asian and Hispanic groups. Aisan male population greater than 6.94 and Asian female population is greater than 7.05 are favored over low-risk class, while Aisan male population less than 5.33 and Asian female population less than 4.34 are favored over high-risk class (in the last example of figure 8).

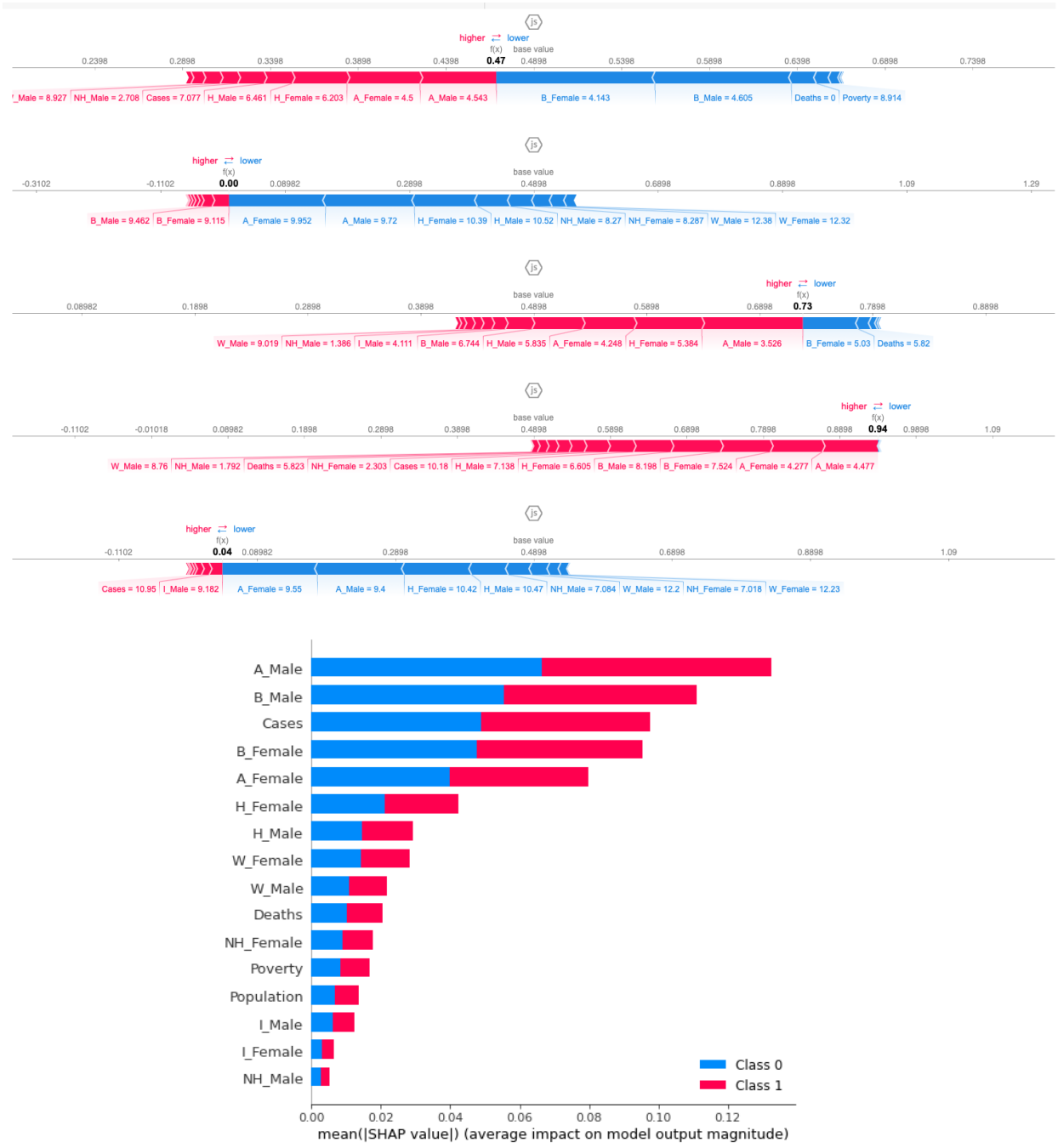

Figure 7: Five Correct predictions explained by SHAP and summary plot

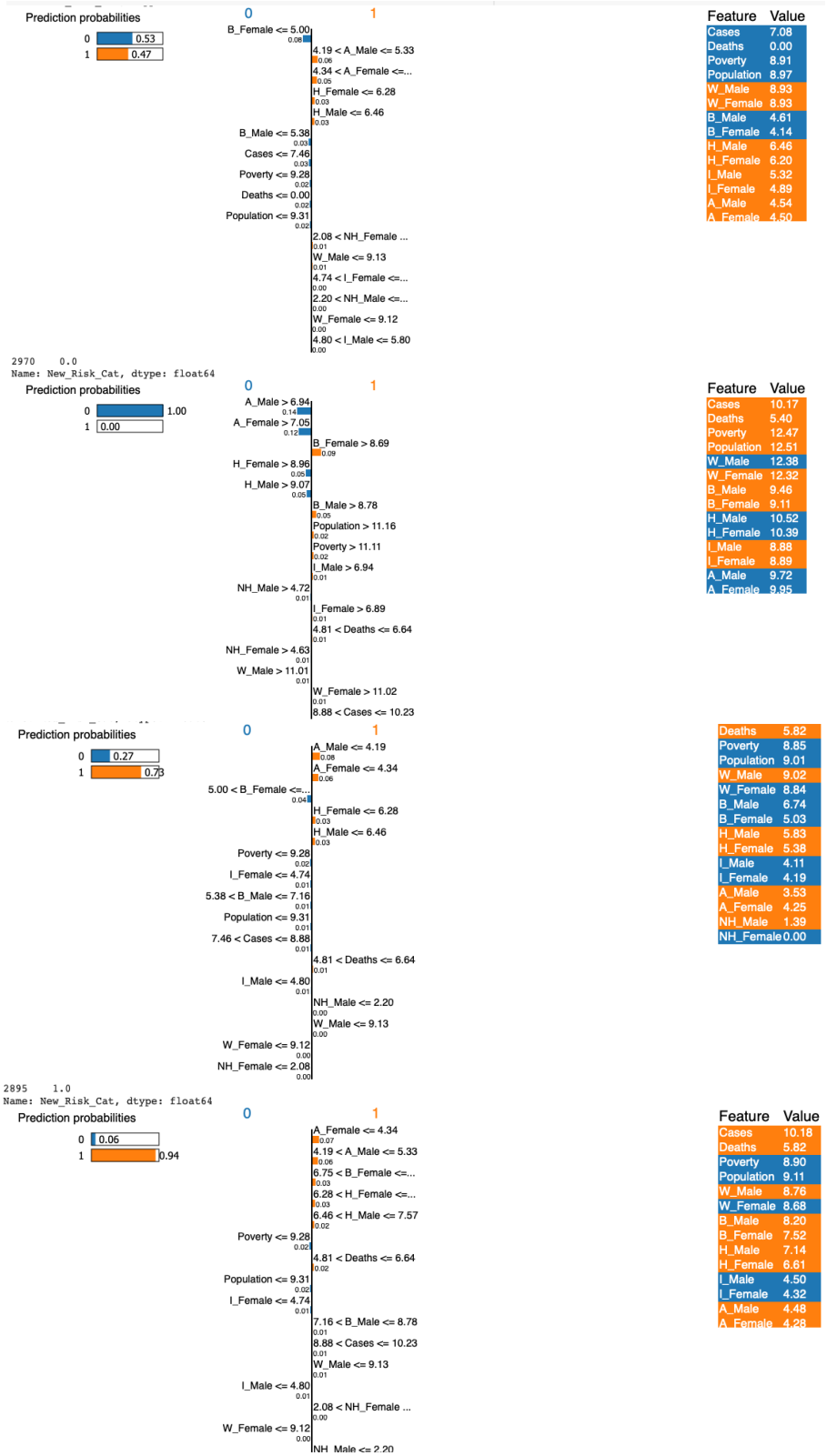

Figure 8: Four LIME explanations of correct predictions

4.2.3 Fairness and demographic parity

According to a published peer-reviewed research on COVID19, individuals from Asian and Black ethnic groups are more likely to be infected by COVID19 compared to those of White ethnicity. The reasonings were that individuals from these minority groups are more likely to live in larger household sizes composed of multiple generations and they tend to have lower socioeconomic status, which may increase the likelihood of living in overcrowded spaces (Shirley Sze et al., 2020).

By trusting the above research findings, fairness here is defined as for all counties, the model should have similar performance in predicting the risk category regardless of the demographic and socio-economic makeup. It turned out that the accuracy in predicting the risk category of counties with more females was 0.77, which is about 0.03 higher than that in predicting counties with more male populations. The FPR are almost similar while the FNR of male population is 0.07 higher. The TPR of more female counties is 0.07 higher (Figure 9).

```
Acccuracy for coutny with more female:  0.7682539682539683
Acccuracy for coutny with more male:  0.732484076433121

 FPR of counties with more male:  0.23776223776223776
 FPR of counties with more female:  0.24305555555555555

 FNR of counties with more male:  0.29239766081871343
 FNR of counties with more female:  0.2222222222222222

 TPR of counties with more male:  0.7076023391812866
 TPR of counties with more female:  0.7777777777777778
```

Figure 9: Fairness metrics of gender

The results of the fairness among racial groups shows that the model had the highest accuracy when predicting counties with more black population, while had the lowest accuracy when predicting counties with more white population.

There's a great difference between the FPR when predicting counties with more Asian or Indian with counties that have more other racial groups. The FPR of more Asian (0.17) and more Indian (0.19) were 0.15 to 0.2 smaller than the FPR of other groups. The highest FPR was observed with counties that have more white populations (0.38). Differently, counties with more had a really high FNR (0.41), followed by Native Hawaiian (0.36), Indian (0.353), and Hispanic (0.324). The lowest FNR was observed with counties that have more black (0.15) populations followed by counties with more white populations (0.21). The highest TPR is associated with counties that have more black population (0.85), followed by counties with more white population (0.8). While the TPR for counties with more other racial groups stays between 0.6 to 0.7 (Figure 10).

```
Accuracy for counties with more White population:  0.7165605095541401
Accuracy for counties with more Black population:  0.7993630573248408
Accuracy for counties with more Hispanic population:  0.7356687898089171
Accuracy for counties with more Asian population:  0.732484076433121
Accuracy for counties with more Indian population:  0.7292993630573248
Accuracy for counties with more Native Hawaiian population:  0.7611464968152867

FPR of counties with more Black:  0.3069306930693069
FPR of counties with more White:  0.38345864661654133
FPR of counties with more Asian:  0.16756756756756758
FPR of counties with more Indian:  0.189873417721519
FPR of counties with more Hispanic:  0.3069306930693069
FPR of counties with more Native Hawaiian:  0.38345864661654133

FNR of counties with more Black:  0.15023474178403756
FNR of counties with more White:  0.20994475138121546
FNR of counties with more Asian:  0.4108527131782946
FNR of counties with more Indian:  0.3525641025641026
FNR of counties with more Hispanic:  0.32432432432432434
FNR of counties with more Native Hawaiian:  0.36496350364963503

TPR of counties with more Black:  0.8497652582159625
TPR of counties with more White:  0.7900552486187845
TPR of counties with more Asian:  0.5891472868217055
TPR of counties with more Indian:  0.6474358974358975
TPR of counties with more Hispanic:  0.6756756756756757
TPR of counties with more Native Hawaiian:  0.635036496350365
```

Figure 10: Fairness metrics among six racial groups

For poverty, our model did better at predicting counties with less poverty populations with an overall accuracy of 0.77. Great difference of FPR, FNR was also observed between counties with less poverty populations and more poverty populations. The FPR for counties with

more poverty was 0.3 higher than that of counties with less poverty. The FNR for counties with less poverty was 0.16 higher than that of counties with more poverty, which also indicates that the TPR of counties with less poverty was 0.16 smaller than that of counties with more poverty (Figure 11).

```
Acccuracy for coutny with more poverty:  0.7261146496815286
Acccuracy for coutny with less poverty:  0.7746031746031746

FPR of counties with more poverty:  0.4112903225806452
FPR of counties with less poverty:  0.11042944785276074

FNR of counties with more poverty:  0.18421052631578946
FNR of counties with less poverty:  0.34868421052631576

TPR of counties with more poverty:  0.8157894736842105
TPR of counties with less poverty:  0.6513157894736842
```

Figure 11: Fairness metrics based on poverty population

5. Summary

5.1 Summary of Kaggle Project

All the public datasets used before the processing are sufficient enough to analyze the COVID risk index, morbidity and mortality of a given county at a basic level. Since the given data was already cleaned and processed with proper distribution among different groups, so the data used was fundamental enough for the Kaggle project and our simple ADS. However, it may not be precise and in-depth enough to reveal a real-word condition of COVID risk level of a county. Considering this, I don't think it's a mature model that is ready-to-use in the public to make deterministic regulations.

The implementation of the Kaggle project is fairly simple in making predictions of morbidity and mortality, which only used a simple random forest regressor with out of bag method to filter out the features' importance. Although not robust, the model still gives a good fit of the data with relatively small R-square and MAE. From the result of feature importance, we can see that black female and poverty ranked as the top two features for morbidity. For mortality (without Case as a feature), black females were the most important feature in determining mortality (deaths count) of a county.

There are two ways to understand the data and result. First, if one believes the original data is not biased (we trust the curator that processed the data) towards any racial groups, the findings stated above indicate that more attention should be paid to the health condition of black females and those who live in poverty during COVID. This benefits the government, the health care system, as well as all US citizens in controlling the spread of COVID19. The government can know which county to focus on based on the demographics makeup, and the healthcare system would know who to prioritize when mitigating COVID trend, and the citizen will benefit greatly if the spread is under control.

The second understanding is that the bias towards black females and people living in poverty already existed in the given data (we do not trust the curator). In this case, it would be helpful to have a rethink of how the morbidity and mortality data was generated. By figuring out where the bias came from, all the stakeholders that are involved in the COVID19 fight will be benefited since this ensures the right resources are used on the right spot.

5.2 Our summary

We used the same data as the Kaggle project (which was also an aggregation of public open data from the government), and we believe the data is explainable for risk categories, but may not be generalizable to the real-world condition.

We first used the random forest classifier to predict the risk category, setting the median risk index as the compared value, with 0 being low-risk and 1 being high risk. We split the data into train and test sets, and fit a random forest classifier to the training set. Then we went on to calculate the accuracy and AUC of the model, which did not turn out as good as we expected. So we tuned the classifier and calculated the statistics again, which showed a significant improvement. Then we used SHAP to explain and analyze the significance of each demographic group, trying to figure if any particular group has more impact than the others.

The observation of fairness metrics based on gender indicates that our model did a better job in predicting counties with more females, and it was more likely to wrongly label high risk counties with more male as low-risk, and did better at correctly predicting high-risk counties with more females. Regarding racial group parity, overall, our model did best in predicting counties with more black populations. However, high FNR was observed with counties that have more Asian or Hispanic populations and high FPR was observed with counties that have more balck or white populations. This indicates that the model was more likely to label low-risk counties with more black or white population as high-risk. The finding that high-risk counties with more Asian are more likely to be labeled as low risk may also explain the research finding that Asian people are more likely to get affected since there might be a chance that these high-risk counties did not receive the proper help due to the prediction is low risk. The model

also did better in predicting counties with less poverty population and was more likely to label a high-risk county with less poverty as low-risk, while more likely to label a low-risk county with more poverty as high-risk.

In conclusion, although the overall and group-wise performance of our ADS (accuracy) was around 0.75, disparity among counties with different distributions of subpopulations groups was observed (especially among balck/white and Asian/Indian). The trends may be picked up from the original data, but we are not sure at this stage since we knew nothing about how the data was combined, generated and preprocessed.

5.3 Limitations

The data used here is in the healthcare field and COVID19 is new to the health care world as well. A lot of features such as age and previous health conditions can also contribute when predicting the risk category/index. So more “private” data needed to be used rather than open data to make insightful and usable predictions. The private data should be protected well and should not be leaked to the public. We are not sure if the original risk was evaluated this way or not, but it may be better to have the risk evaluated at an individual level and then at a county level, which could provide more insightful and suggestive results.

For future ADS, larger and more detailed COVID data is needed, and the raw data should be transformed to protect privacy, which can be done in a black box model. While protecting privacy, the ADS should also have different degrees of transparency to different stakeholders so they can trust and use the result generated by the ADS. If the ADS will be put into use to male real-world decisions, the AUC should be at least 0.9 or even 0.95 to make effective and proper recommendations. Also the ADS should also take into account the fact that different population

groups may truly have different levels of risk in getting affected (based on current research), which means it is reasonable to put different weights to different population groups when training the model. However, although different weights, the fairness metrics such as FPR and FNR should remain similar among different groups.

In the future, we can also evaluate the model's performance on counties with more population of two of the demographic features. For example, we can compare performance in counties with more Black and White populations with counties that have more Asian and Indian populations. Thus, we may be able to observe if a combination of features may have a higher impact on the prediction outcome.

6 Notes about the code

For the last parts of the codes that were related to fairness calculation, we accidentally added new columns to the input test data for analysis, which should be prohibited. Due to the time constraint and the fact that the model predicts the test data before the test set was changed, we didn't change the code for later parts. We should pay more attention in the future coding process.

References:

Laurindogarcia. “ML: Covid 19, Race, Gender, Poverty, Health.” *Kaggle*, 26 Sept. 2020, www.kaggle.com/laurindogarcia/ml-covid-19-race-gender-poverty-health.

Narkhede, Sarang. “Understanding AUC - ROC Curve - Towards Data Science.” *Medium*, 14 Jan. 2021, towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

Sze, Shirley. “Ethnicity and Clinical Outcomes in COVID-19: A Systematic Review and Meta-Analysis.” *Thelancet.Com*, 12 Dec. 2020, www.thelancet.com/journals/eclinm/article/PIIS2589https://www.thelancet.com/action/showPdf?pii=S2589-5370%2820%2930374-6-5370(20)30374-6/fulltext#secccesectitle0021.