**Predicting the Severity of Incoming COVID-19 Cases in Filipino Health Networks**   
**DS-GA 1003 Final Project**

Lauren D’Arinzo (lhd258), Elizabeth Combs (eac721), Angela Marie Teng (at2507) - *responsible*, Paula Kiatkamolwong (kk4158), Steven Dornberg (dornbs01)

1. **Introduction**

In 2020, the worldwide COVID-19 pandemic continues to upend the lives of billions. Despite the influx of news coverage about the coronavirus health crisis, it is evident that access to public health information may not be available to everyone. National news outlets in the Philippines have been shut down, and access to accurate and timely data about COVID-19 is becoming more and more sporadic. In the midst of national shutdowns, other risks have become apparent: political unrest and economic turmoil. Our project discusses and evaluates the current situation of COVID-19 in the Philippines, using datasets sourced from the Philippine Department of Health.

We sought to implement a supervised model which, given a vector of input features from historical COVID-19 cases in the Philippines, could predict whether a patient's outcome would be mild or severe. If successfully trained and deployed, this type of model could allow health systems and policymakers to better anticipate the individual needs of practices and hospitals and better allocate resources based on data-driven research.

We focus on Philippine data from the national Department of Health because of its patient-level reporting. Unlike other available datasets, which are typically aggregated at a country-level and which may not have information about patient comorbidities, patient-level dataset allows us to answer the research question: Given a presumed COVID-19 diagnosis and features including age, travel history, epidemiological virus links, can we predict a person’s health status in relation to COVID-19? In particular, will they have a mild or severe case? Answering this question will allow Filipinos to make more informed personal and community healthcare decisions, as well as provide government officials with statistics that will give them the additional tools to combat this pandemic.

1. **Related Work**

As the country with the third highest number of cases in Southeast Asia, the Philippines has been on an enhanced community quarantine since March 12. Work has been done to elucidate the global spread and transmission of SARS-CoV-2 (COVID-19, novel coronavirus), although only a number of studies have been published in the Philippine context. Edrada et al. gave an overview of the spread of the virus across the archipelago, tracing the roots of the first two cases in the Philippines to “previously healthy Chinese nationals on vacation” in January 2020[[1]](#footnote-0). Additionally, Rabajante et al. used mathematical models to show that exposure time is a “significant” factor in the spread of the disease[[2]](#footnote-1). Current literature can be better supported by supervised learning methods for prediction.

1. **Approach**

***Dataset Description***

The [dataset](https://docs.google.com/spreadsheets/d/16g_PUxKYMC0XjeEKF6FPUBq2-pFgmTkHoj5lbVrGLhE/edit?fbclid=IwAR29_EijSX1_bTS1oNW6nuYNjMvoN-HcP44PCEoM13JSfXD0gf1Uu2D2XoE#gid=0) used for this project comes from the Philippine Department of Health (DOH). The spreadsheet is a live source gathered by Filipino data scientists and obtained from the DOH records of reported and publicly announced[[3]](#footnote-2) COVID-19 cases. The dataset was compiled and is currently managed by the Philippine Data Science [Group](https://www.facebook.com/datasciencephilippines). We chose this dataset because it contains [patient-level data](https://public.tableau.com/profile/rozauro.cordon?fbclid=IwAR1EaH7YlwhyylXkE4reI3wwBlbayTcyrl1W3KbxDLTnRnHo5BU-Mc_6P0I#!/vizhome/COVID-19PHTracker/COVID-19?publish=yes)[[4]](#footnote-3); all other publicly available datasets related to clinical COVID-19 observations that we were able to find only contain aggregate counts by country or by city.

*Limitations of Dataset Scope*

However, because COVID-19 is an ongoing challenge, the DOH dataset is constantly being updated. Since the prioritization in most health networks right now is efficient care, data documentation is lagged. We were restricted to a total sample size of 157 complete instances[[5]](#footnote-4), as many COVID cases since April are documented but comorbidities and other details of the case remain ‘for validation.’ Additionally, the DOH is notorious inaccurately collecting data, and case underreporting and lack of testing is a national issue. Many of the COVID-19 tests have been reserved for national officials, or wealthier and socially well-off members of the country. Thus, although we have a diverse sample of patients, we are uncertain whether or not it represents the population of Philippine patients who have tested positive for COVD-19.

***Data Preparation***

*Target Variable*

The original target variable of interest was a multi-class variable ‘Health Status’, which took on increasing values of severity: recovered, asymptomatic, mild, severe, critical, and died. Given our small sample size that resulted in even smaller within-class sample sizes, we decided to see if there were any naturally-occurring groupings within the categorical response variable[[6]](#footnote-5). We grouped the response variable into two groups: (1) recovered, asymptomatic, and mild (2) severe, critical, and died. This helped reduce the uneven distribution of categories leaving 85 “mild” and 72 “severe” cases, and gave us the opportunity to think about how our project could be useful in a business setting--after research, we discovered that predicting the general severity of a patient (mild vs severe) would be more useful in the Philippine context today, compared to predicting the specific gravity of a case.

*Feature Engineering*

Because the dataset is a live document, it required extensive data cleaning and feature engineering to be compatible with supervised learning approaches. The dataset had both numeric and categorical variables. For the numeric variables, we used normalization to scale the data and minimize bias, as we were dealing with feature values that ranged from 10 (age) to 1,000,000 (population). To handle the categorical variables, we used dummy variables to create a sparse dataframe, which could then be used for modeling. Prior to one hot encoding, we used OpenRefine[[7]](#footnote-6) to correct any misspellings and manual-entry errors. We used different text clustering methods[[8]](#footnote-7) to combine symptoms, comorbidities, other diseases, and locations to ensure the accuracy of our dataset.

*Symptoms and Comorbidities*

When we apply one-hot encoding, each symptom and previous medical history ailment will become its own binary column. Due to the nature of free text input, medical terminology and grammar mistakes are prevalent in this kind of dataset. In order to get the most accurate results on feature importance, it is imperative that there are not various columns that are overlapping and/or duplicates. An example of obvious overlap that could lead to inaccurate results and incorrect weighing of feature importance would be having two different one-hot encoded columns that are for the same thing, such as dis\_renal\_disease and dis\_kidney\_disease, or dis\_cardiovascular\_disease and dis\_heart\_disease. In order to remedy this, data cleaning and grouping was heavily applied in respect to normalizing and spelling. We have went through all symptoms and previous diseases, and grouped and renamed ones that are part of the same primary condition(diabetes, heart disease, renal disease), corrected misspellings, and characterized similar ailments under their correct parent hierarchy (For instance, mapping cardiac disease and coronary artery disease both under cardiovascular disease). Data cleaning in these types of situations will aid in the models accuracy and is a necessary part of working with free-text data.

*Epidemiological Link*

The original dataset contained a feature called ‘Epi Link,’ which was a free text field describing any exposure that patient had to other COVID-19 patients. For example, for a given patient, the field might contain “wife of ph42, contact with other patient.” We attempted to quantify this exposure by counting the number of exposures using regular expressions (regex[[9]](#footnote-8)). The given example would then have an Epi\_Link\_Exposure value of 2.

*Date Variables and Data Leakage Concerns*

There were multiple date fields in the original dataset, such as Date of Onset of Symptoms, Date of Admission, and Date Final Status Was Documented. We originally derived numeric features representing the various permutations of differences in these variables, for example, ‘Days between hospital admission and Date of final status.’ However, we were hesitant to include these date features as modeling inputs because of the potential data leakage implications; in a deployed version of this model, a hospital would like to predict the outcome of a patient when they are admitted. In this scenario, the hospital would not know the ‘Date of final status,’ because it has not happened yet. Consequently, we decided not to use these features explicitly for modeling, but we did explore their distributions in our exploratory data analysis,[[10]](#footnote-9) considering that the conclusion of Rabajante et al. described how exposure time plays a substantial role in the transmission of COVID-19. Unfortunately, we did not have access to features that reflected the time of exposure and onset of symptoms.

*Geolocation*

In the original dataset, there were two main location columns; the city\_name, which contained the city location of the patient, and the long\_lat, which contained the longitude and latitude of the patient’s address. Upon inspection, we discovered that the city column contained potentially misleading and inaccurate information. Various cities in the same country existed in different provinces, moreover, locations in the city column sometimes contained the city, and sometimes contained the municipality. Thus, we utilized the longitude and latitude information instead. Using GeoPy[[11]](#footnote-10) and GeoPandas[[12]](#footnote-11), we reverse geo-encoded each entry, with a resulting output of a json-object that contained standardized location data. From there, we extracted each city and region to obtain each location’s corresponding population data.

*Population-level Census Data*

Because the dataset we used contained geo-location and patient-level data, we were able to extract additional features related to this information. First, we derived a new feature called average\_income, where the value for each entry corresponds with the average income for a particular region that the patient lives. To obtain each city’s average income, we used the Philippine Statistics Authority’s Census Data[[13]](#footnote-12). Second, we extracted another variable called population\_density where we divided each patient’s location into different geographic levels, like regions and cities, and then mapped each patient to their city’s population density accordingly.

*Missing Value Treatment*

Most of the missing values in our dataset occurred in the target variable--because COVID-19 takes about 14 days to incubate, many of the patients in our dataset of ~4000 cases were still hospitalized. We decided to drop the rows which correspond to patients who are still undergoing treatment, as it wasn’t clear whether they had a mild or severe case of COVID-19.

*Train/Validation/Test Split*

We randomly split our data into 60% train, 20% validation, and 20% test sets for training and experimentation across tree-based and linear algorithms. Typically, we would set the training set as the earliest cases, the validation set as the cases that occured after training, and the test set as the future cases, since the main goal of our project would be to predict COVID-19 case severity given medical and population data. However, given our small total sample size of 157 patients[[14]](#footnote-13), our team decided to do a random split across all our samples. Although our dataset may be influenced by time, we decided to apply a random split because of our dataset constraints, as well as our prior assumption that each case was independent, and that the underlying distribution of this specific population was identically-distributed.

1. **Experiments:**

***Feature Selection***

To determine which features contribute the most to our target variable, we ran a feature importance algorithm on the gradient boosted model. After eight features, the dropoff in variance seems marginal, which prompted us to retrain a model using only the top eight most important features. The decrease in accuracy between a model that contains all 85 features compared to a model that contained only 8 features was 4%. This accuracy decrease is small enough that we decided to continue hyperparameter-tuning on the 8-feature model, due to time and efficiency constraints.

Since feature importance is only applicable to linear support vector classifiers, we used the relative weight values of each feature instead. We applied dimensionality reduction using an L1 penalty, and the primal form of the support vector classifier. Using the SelectFromModel() and transform() functions, we created a new dataframe with 30 features instead of our original 88. Because the model accuracy only decreased by 2%, we decided to use these new dimensionality-reduced feature space for hyperparameter tuning.

***Results of Baselines and other approaches***

Overall, the best performing classifiers were the hypertuned gradient boosted model, and linear support vector classifiers (Figures 1, 4, 7)

***Hyperparameter tuning***

To tune the gradient boosting model, we used GridSearch and 3-fold cross-validation. We found the best parameters to be: {'learning\_rate': 0.1, 'max\_depth': 1, 'min\_samples\_leaf': 1, 'min\_samples\_split': 30, 'n\_estimators': 30,’ 'random\_state': 42}. To select the optimal number of n\_estimators, we plotted the training and validation accuracy over a range of n\_estimators, to observe the change in accuracy as we increased the number of boosting stages performed. After 30 boosting stages, we see only incremental increases in accuracy, which prompted our team to select the optimal n\_estimators = 30 in response to efficiency constraints. Overall, our best gradient boosting model achieved a test accuracy of 72%.

To tune the support vector classifier, we also used GridSearch and 3-fold cross-validation. We found the best parameters to be: {‘C’: 1, ‘max\_iter’: 10, ‘random\_state’: 42}. To select the optimal regularization parameter, we created a plot of the training and validation accuracy over different values of C. Overall, our best support vector classifier achieved a test accuracy of 77%.

***Error analysis***

*Evaluation Metrics*

The primary metric we used for evaluation was accuracy, which captures the proportion of all correctly identified cases. Additionally, we also calculated F1 score. As the harmonic mean of precision and recall, the F1 score provides additional evaluation of the incorrectly classified cases. This is particularly important in clinical health applications, because the implications of false negatives are often worse than false positives. In the case of our classifier, a false positive is an instance where a COVID-19 case is predicted to be severe, but is actually mild, and a false negative is an instance where the case is predicted to be mild but is actually severe.

***Explanation of Results***

*Linear Support Vector Classifiers*

Support vector machines (SVC)[[15]](#footnote-14) are linear classifiers that construct separating hyperplanes in high dimensional spaces, and can be used for classification or regression tasks[[16]](#footnote-15). Essentially, SVC maximizes the margin by minimizing the L2 norm of the weights vector, subject to a penalty given the corresponding constraint. SVC may work particularly well with our dataset because it is a binary classification problem, and linear kernels may be optimized in this scenario, given that the data distribution is independent and identically distributed. We know that our classes have a linear boundary because the SVC[[17]](#footnote-16) with a linear kernel outperformed those with a radial basis function, polynomial, and sigmoid kernel[[18]](#footnote-17). Because SVC uses hinge loss to maximize the margin, we can see that our dataset may be linearly separable with a hyperplane. At penalty level 1, the “inverse regularization parameter” gives the highest test accuracy.

*Gradient Boosting Model (XGBoost)*

Gradient boosting is an “additive model in a forward stage-wise fashion”[[19]](#footnote-18), and is useful for modeling non-linear prediction problems. In particular, because gradient boosting is a tree-based model, it is generally effective when dealing with datasets that have a lot of categorical features. The parameter n\_classes\_ determines the number of trees that are fit on the “negative gradient of the binomial or multinomial deviance loss function”[[20]](#footnote-19). In the case of binary classification, n\_classses\_ would refer to a single regression tree. By taking an ensemble of weak learners, the XGBoost algorithm optimizes a given cost function by iteratively choosing an approximation that minimizes the expected value of the loss function[[21]](#footnote-20). Like random forests, gradient boosting uses a set of decision trees--although instead of using the majority rules approach, gradient boosting models build decision trees one at a time, and improves on existing weak learners[[22]](#footnote-21). This additive approach may be one reason why gradient boosting outperformed random forests in our dataset, and achieved the highest test accuracy out of all the tree-based models.

**Discussion:**

*Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? If you were to continue working on the project, what are the interesting areas for future work? What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.*

*Evaluation of Findings (two best models)*

*[Paula and Steven]*

*Possible Next Steps (deployment → more data, real-time) target variable → multi-class ensemble methods*

*[Paula and Steven]*

**Conclusion:**

**Appendix:**

All code and data for this project is available at <https://github.com/angelaaaateng/ML_COVID_PREDICTION>

Figure 1. Results of baseline and other approaches

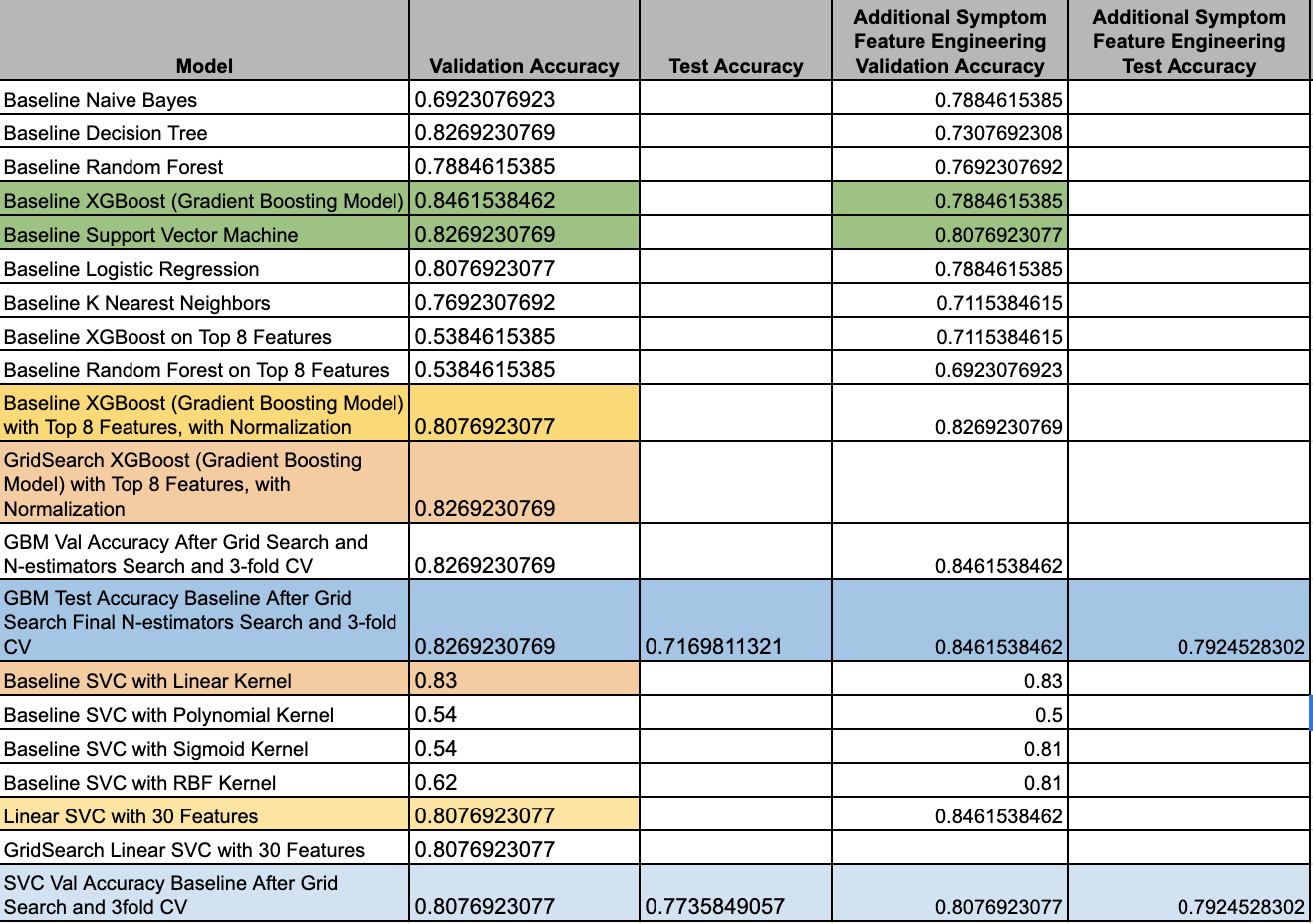


Figure 2. Results of K-Means Clustering

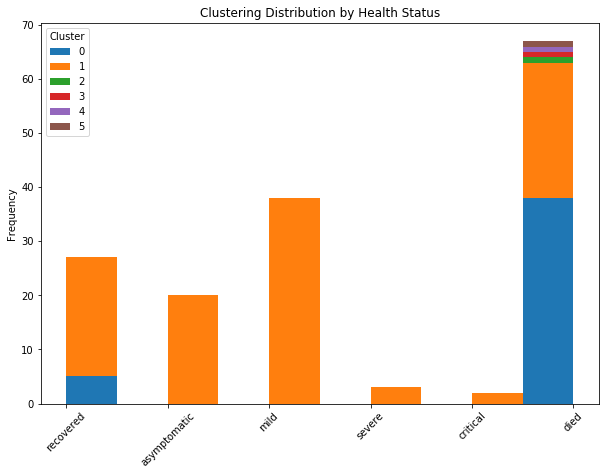
****

Figure 3 .Best Performing Tree-Based Model: Gradient Boosting Model

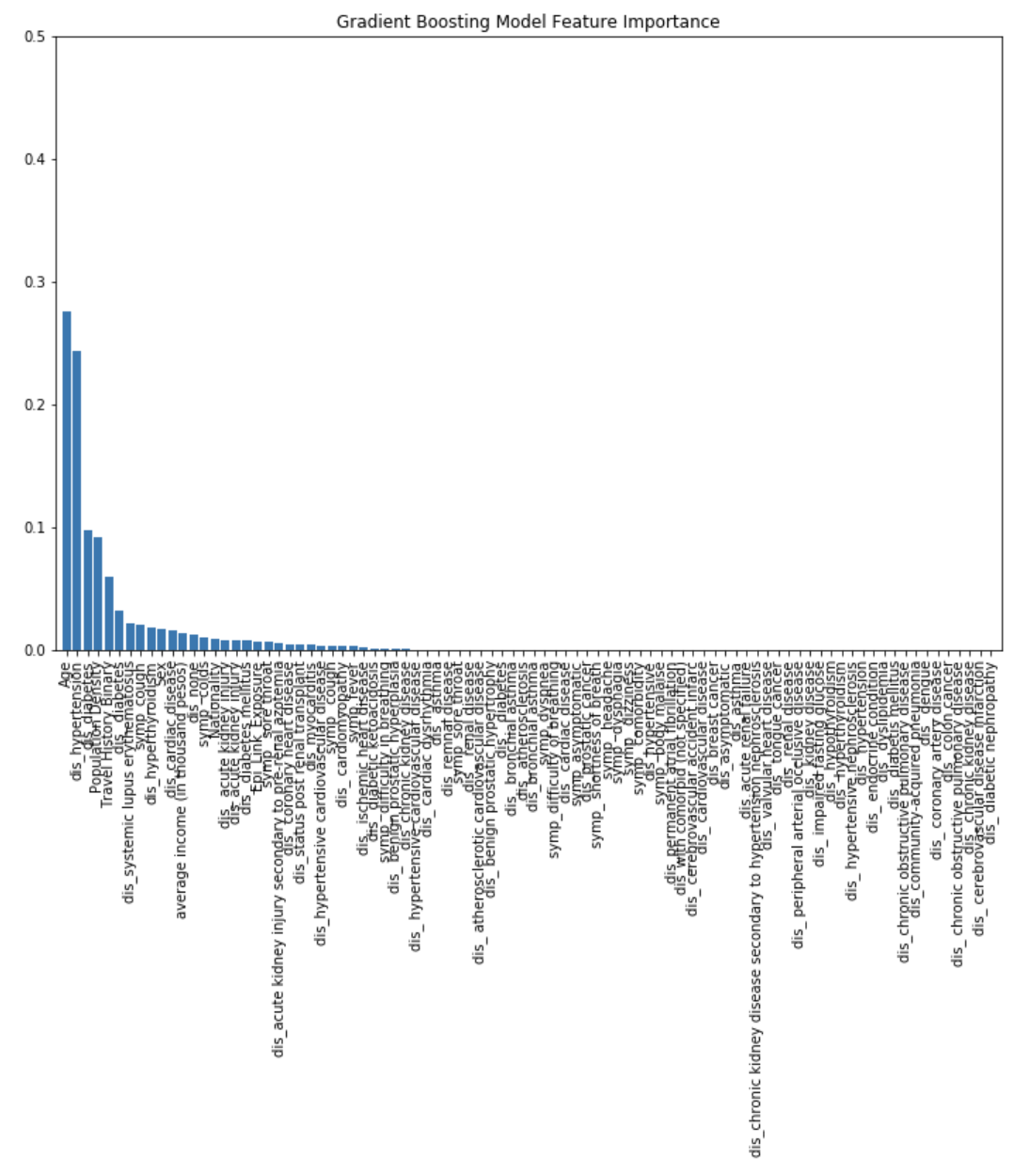
****

Figure 4. Result of Best Performing Linear Model, Support Vector Machine, on Validation Set

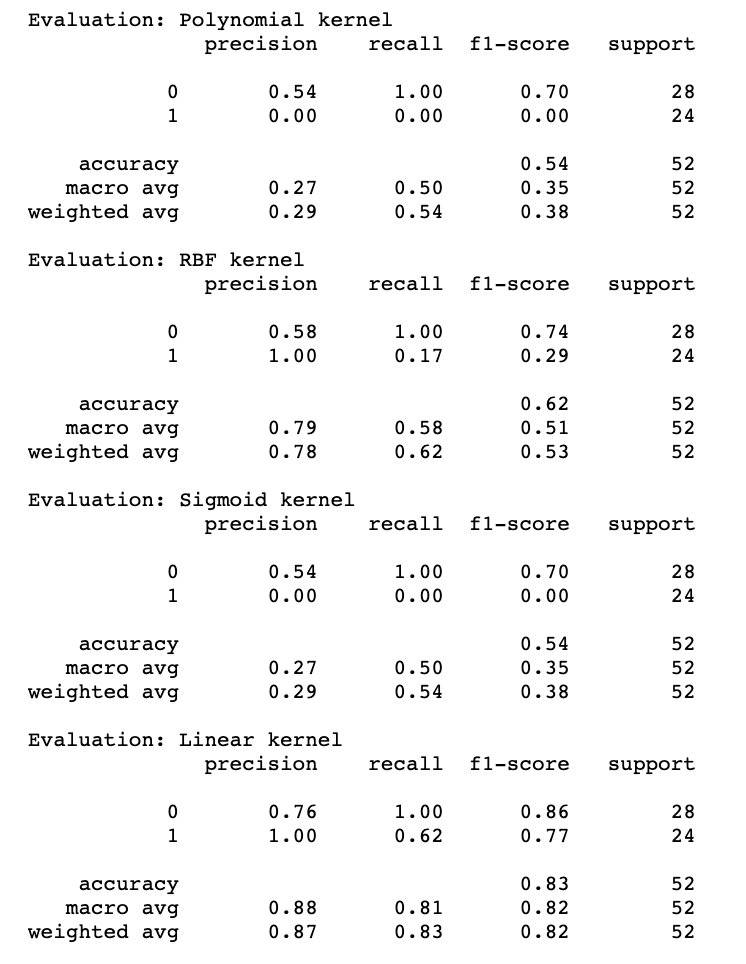
****

Figure 5. Results of Hyperparameter Tuning, N-Estimators Parameter

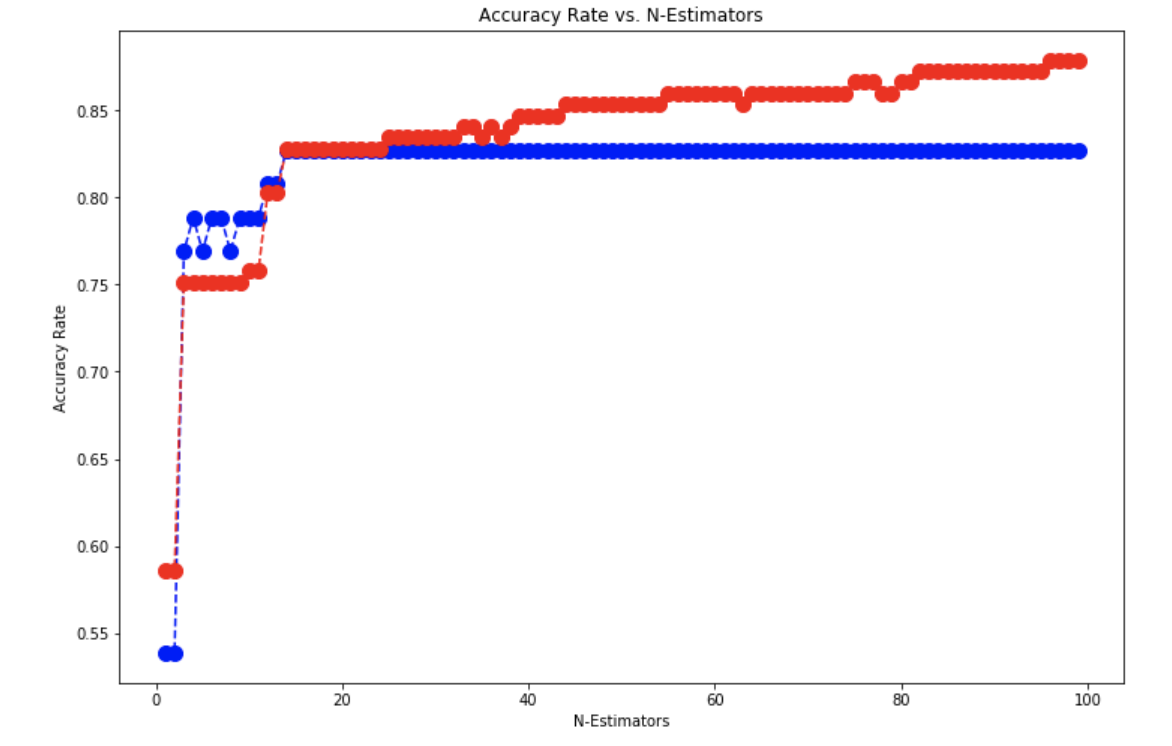
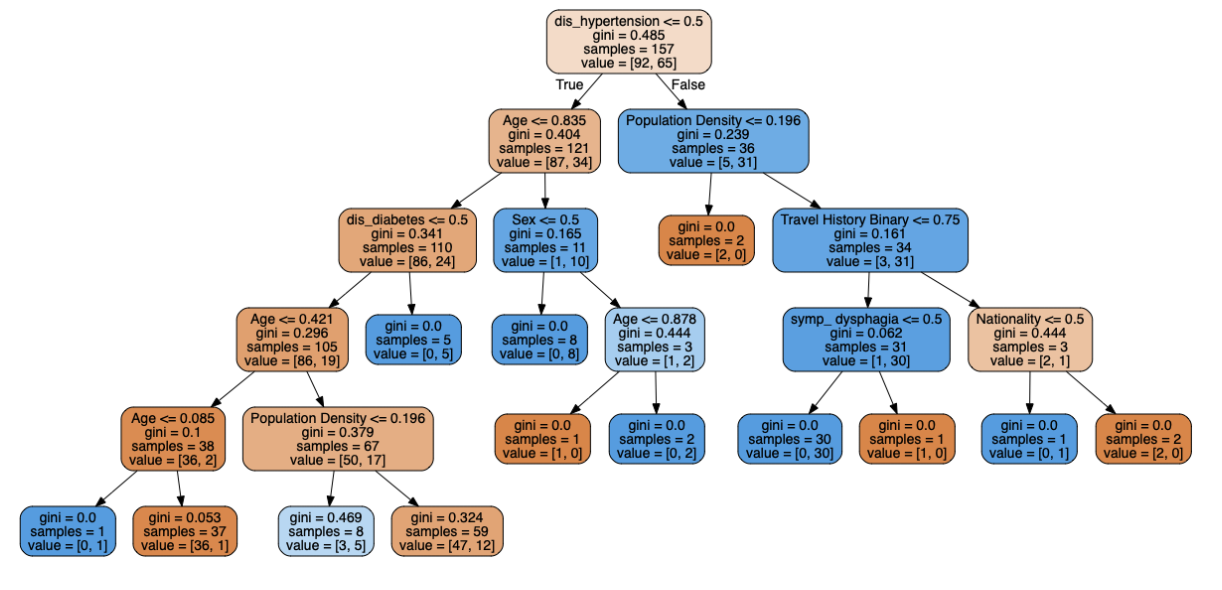
****

Figure 6. Feature Importance Results of Decision Tree at Max Depth 5 (A) and Max Depth 3 (B)

****

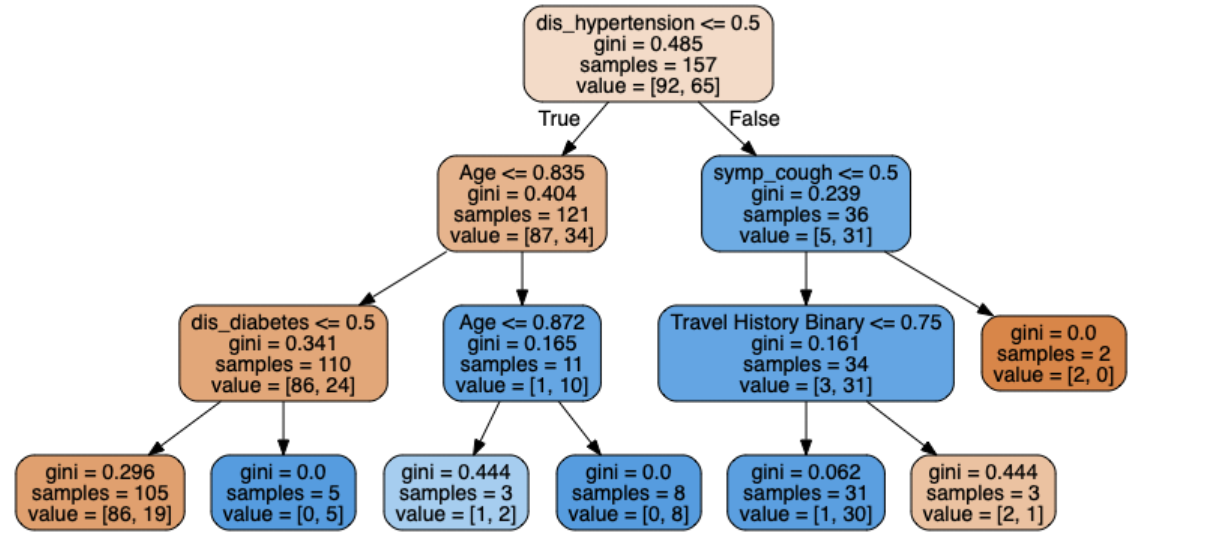
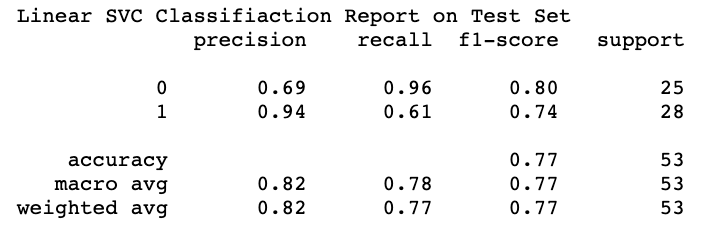
****

Figure 7. Result of Best Performing Linear Model, Support Vector Machine, on Test Set

****

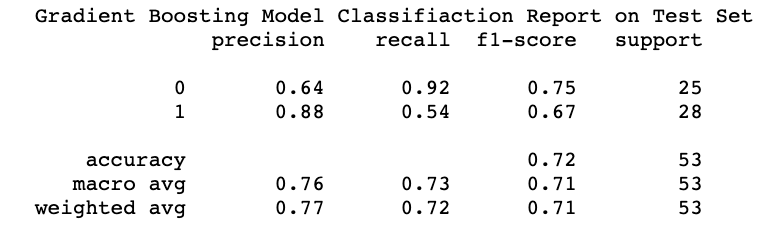
****

Figure 8. Pairwise correlations of numeric features

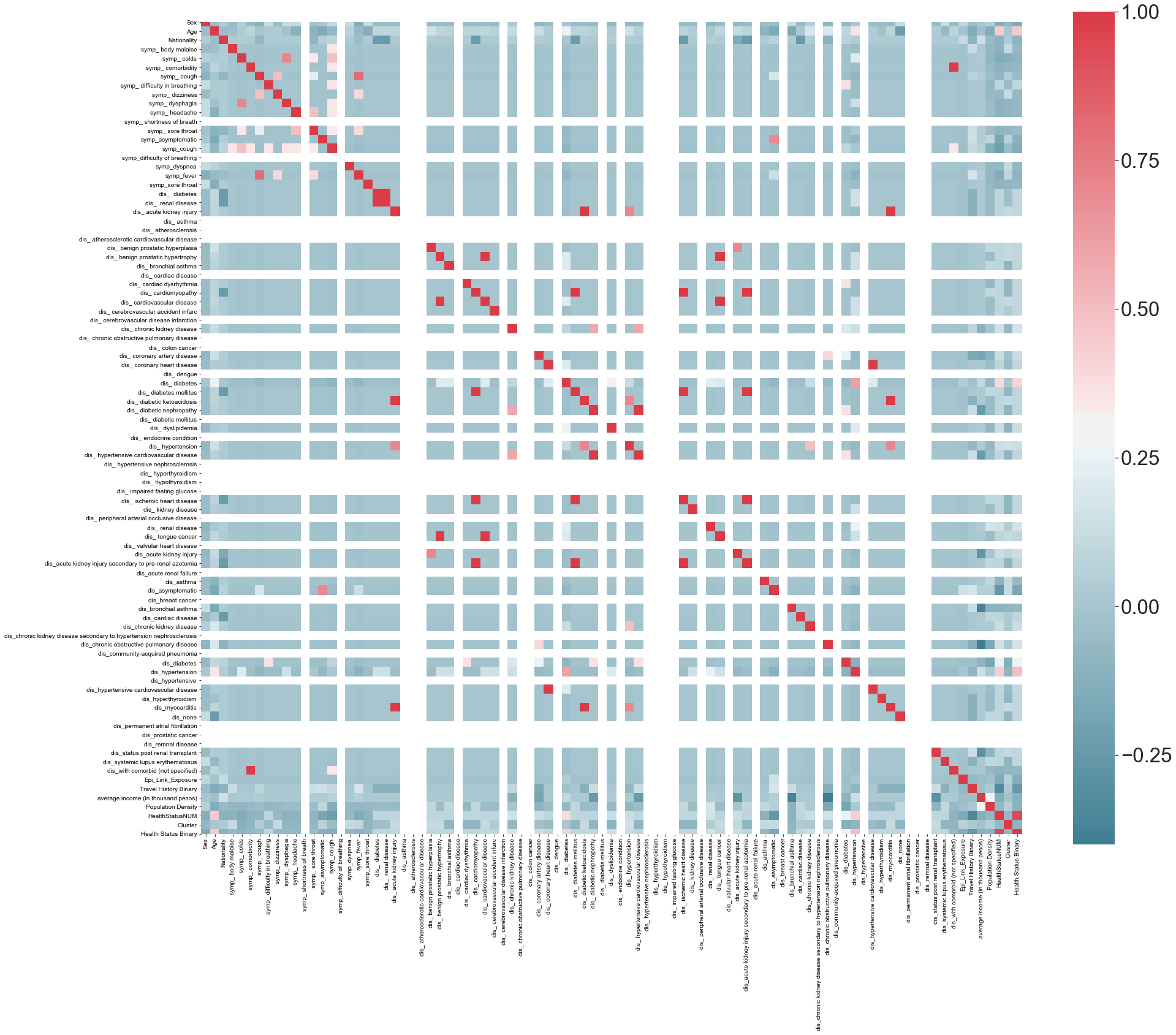
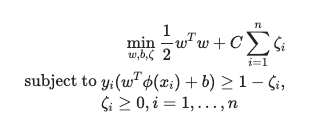
****

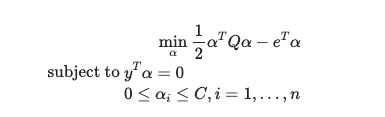
Figure 9. Objective Functions and Primal/Dual Forms:

SVC

The primal problem that SVC[[23]](#footnote-22) solves is the following:



And the dual problem it solves is[[24]](#footnote-23):



**Bibliography:**

Chen, Nanshan, et al. “Epidemiological and Clinical Characteristics of 99 Cases of 2019 Novel Coronavirus Pneumonia in Wuhan, China: A Descriptive Study.” *The Lancet*, vol. 395, no. 10223, Lancet Publi

Edrada, Edna M., et al. “First COVID-19 Infections in the Philippines: A Case Report.” *Tropical Medicine and Health*, vol. 48, no. 1, BioMed Central Ltd., Apr. 2020, doi:10.1186/s41182-020-00203-0.

Kay, Cherish, and L. Pastor. “Sentiment Analysis on Synchronous Online Delivery of Instruction Due to Extreme Community Quarantine in the Philippines Caused by COVID-19 Pandemic.” *Asian Journal of Multidisciplinary Studies*, vol. 3, no. 1, 2020, https://asianjournal.org/online/index.php/ajms/article/view/207.

Kong, Weifang, and Prachi P. Agarwal. “Chest Imaging Appearance of COVID-19 Infection.” *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, Radiological Society of North America (RSNA), Jan. 2020, p. e200028, doi:10.1148/ryct.2020200028.

Nicomedes, CJ, and R. Avila. *An Analysis on the Panic of Filipinos During COVID-19 Pandemic in the Philippines*. 2020, doi:10.13140/RG.2.2.17355.54565.

Rabajante, Jomar F. *Insights from Early Mathematical Models of 2019-NCoV Acute Respiratory Disease (COVID-19) Dynamics*. Feb. 2020, http://arxiv.org/abs/2002.05296.

Schultz, Marcus J., et al. “Current Challenges in the Management of Sepsis in ICUs in Resource-Poor Settings and Suggestions for the Future.” *Intensive Care Medicine*, vol. 43, no. 5, Springer Verlag,

Sohrabi, C., et al. “World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19).” *Elsevier*, https://www.sciencedirect.com/science/article/pii/S1743919120301977. Accessed 15 May 2020.

Sohrabi, Catrin, et al. “World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19).” *International Journal of Surgery*, vol. 76, Elsevier Ltd, 1 Apr. 2020, pp. 71–76, doi:10.1016/j.ijsu.2020.02.034.

Watanabe, Shumpei, et al. “Bat Coronaviruses and Experimental Infection of Bats, the Philippines.” *Emerging Infectious Diseases*, vol. 16, no. 8, 2010, pp. 1217–23, doi:10.3201/eid1608.100208.

Wenham, C., et al. “COVID-19: The Gendered Impacts of the Outbreak.” *Thelancet.Com*, https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30526-2/fulltext?te=1&nl=in-her words&emc=edit\_gn\_20200317. Accessed 15 May 2020.

Wu, YC, et al. “The Outbreak of COVID-19: An Overview.” *Ncbi.Nlm.Nih.Gov*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153464/. Accessed 15 May 2020.

Wu, Yi Chi, et al. “The Outbreak of COVID-19: An Overview.” *Journal of the Chinese Medical Association*, vol. 83, no. 3, Wolters Kluwer Health, 2020, pp. 217–20, doi:10.1097/JCMA.0000000000000270.

1. Edrada, Edna M., et al. “First COVID-19 Infections in the Philippines: A Case Report.” Tropical Medicine and Health, vol. 48, no. 1, BioMed Central Ltd., Apr. 2020, doi:10.1186/s41182-020-00203-0. [↑](#footnote-ref-0)
2. Rabajante, Jomar F. Insights from Early Mathematical Models of 2019-NCoV Acute Respiratory Disease (COVID-19) Dynamics. Feb. 2020, http://arxiv.org/abs/2002.05296. [↑](#footnote-ref-1)
3. These cases identify patients by PH\_id, and are publicized in national newspapers like The Inquirer [↑](#footnote-ref-2)
4. Each case is at an individual patient level [↑](#footnote-ref-3)
5. Where each complete instance is defined as a case where we know the patient outcome (no pending cases) [↑](#footnote-ref-4)
6. Was the problem a true multi-class problem or would it be most important to predict mild or severe cases? Using the k-means++ algorithm on standardized training features, we were able to group the multi-class health status into a binary one by mapping the clusters back to the response variable and examining the distributions. This finding corroborated our business understanding that individuals and practitioners would be most interested in a prediction of whether a COVID case would become severe or not. [↑](#footnote-ref-5)
7. <https://openrefine.org/> [↑](#footnote-ref-6)
8. For the full process and the various permutations we used to ensure our data was clean, please see the saved OpenRefine project and documentation on our github repo here: <https://github.com/angelaaaateng/ML_COVID_PREDICTION> [↑](#footnote-ref-7)
9. <https://regex101.com/> [↑](#footnote-ref-8)
10. Please see our master.ipynb on our Github for the full EDA [↑](#footnote-ref-9)
11. <https://geopy.readthedocs.io/en/stable/> [↑](#footnote-ref-10)
12. <https://geopandas.org/geocoding.html> [↑](#footnote-ref-11)
13. The CSV files for the raw income data can be obtained here <http://www.psa.gov.ph/> [↑](#footnote-ref-12)
14. The response variable distribution varied significantly over time since new, active cases are appended while past cases' health statuses are finalized. The data generation process is another reason we felt it most accurate to group the data into a binary response. [↑](#footnote-ref-13)
15. <https://scikit-learn.org/stable/modules/svm.html#svm-kernels> [↑](#footnote-ref-14)
16. For more information on SVC, and their primal/dual forms, please see the appendix [↑](#footnote-ref-15)
17. <https://scikit-learn.org/stable/modules/svm.html#svm-kernels> [↑](#footnote-ref-16)
18. <https://scikit-learn.org/stable/modules/svm.html#classification> [↑](#footnote-ref-17)
19. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html> [↑](#footnote-ref-18)
20. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html> [↑](#footnote-ref-19)
21. <https://en.wikipedia.org/wiki/Gradient_boosting> [↑](#footnote-ref-20)
22. <https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained> [↑](#footnote-ref-21)
23. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> [↑](#footnote-ref-22)
24. <https://scikit-learn.org/stable/modules/svm.html#svm-kernels> [↑](#footnote-ref-23)