***Comparative Analysis of COVID-19 Mortality Rate and Diet***

***Group Project***

**I.INTRODUCTION**

The 2019 novel coronavirus (SARS-CoV-2) is a new highly contagious virus that emerged in Wuhan, China causing a worldwide pandemic. With millions of globally confirmed cases and deaths, COVID-19 has caused an unprecedented amount of change in people's lives. Because of this, people have been prioritizing their health now more than ever. This study focuses on analysis of datasets to uncover the severity of the viruses' effects based on diet. We will analyze the metabolic rates of people from countries and their COVID-19 mortality rate. This study is being conducted in an effort to understand the virus and help people around the world combat it. Depending on our findings, we can then use this information to adjust our own diets accordingly.

We plan to find and analyse the predictive variables that show associations and patterns to prove our hypothesis. Metabolic health is defined as having ideal levels of blood sugar, triglycerides, high-density lipoprotein (HDL) cholesterol, blood pressure, and waist circumference, without using medications. These factors are directly related to heart disease, diabetes, and strokes. This study focuses on the comparative analysis of the COVID-19 mortality rate and diet because our bodies perform metabolic processes to make or get energy from foods we eat.

Different foods are composed of different compounds such as proteins, carbohydrates, and fats. Our bodies are fueled when our digestive system uses enzymes to break down the food into sugars and acids. Our body can either store the energy it creates from the foods in our body tissues like liver, muscles, and body fat, or it can use the fuel right away. This is why we need a balanced diet. The Center for Nutrition Policy and Promotion recommends a diet intake guideline of 30% grains, 40% vegetables, 10% fruits, and 20% protein.

Metabolic disorders are grouped into categories. These groups depend on the type of abnormal chemical reactions that happen in our bodies because we either lack or have too high of an intake of certain foods. For example, d[iabetes](https://medlineplus.gov/diabetes.html) is a metabolic disorder that causes the diseased or abnormal functioning of organs. Another group is one that affects the breakdown of [amino acids](https://medlineplus.gov/aminoacidmetabolismdisorders.html), [carbohydrates](https://medlineplus.gov/carbohydratemetabolismdisorders.html), or [lipids](https://medlineplus.gov/lipidmetabolismdisorders.html). And lastly, a group where cells part that produce energy are affected, like [mitochondrial diseases](https://medlineplus.gov/mitochondrialdiseases.html). The treatments that can correct symptoms of metabolic disorders focus on diet corrective approaches. The objective of this study is to promote corrective and preventive measures, a crucial step in an effort to save lives against the COVID-19 virus. From the dataset, we can gather information regarding diet patterns from countries with lower COVID infection rate, and adjust our own diet accordingly.

**II. RELATED WORK**

COVID-19 has impacted our lives through various aspects, even taking the lives away from many. Researchers and scientists have been working relentlessly, trying to figure out some health issues which may cause higher mortality rates after being affected by COVID-19. Quoted from a research study conducted by Joseph E. Ebinger along with other researchers, it was said that, “Deceased patients were more likely to have elevated markers of inflammation. Baseline age, BMI, oxygen saturation, respiratory rate, WBC count, creatinine, and ALT were significant prognostic indicators of mortality. Diet is a factor of all the factors in bold” [1]. It also stated that they utilized multivariable logistic regression to find out the “associations of pre-existing traits with a Covid-19 illness severity defined by level of required care” [1]. Other studies reinforced the effectiveness of using logistic regression to examine the associations [6]. The result indicated that patients who are obese, with diabetes mellitus, and with high blood pressure, which are caused by the disorder on diet, along with other variables suffer greater COVID-19 illness severity [1].

Other health factors like cytokine storms and obesity should be taken into consideration. In a research paper written by Tao Liu and other researchers, it was indicated that “seline IL-6 was also significantly related to the increase of baseline level of CRP, LDH, ferritin and D-dimer. The increase of baseline IL-6 level suggests that it may positively correlate with the severity of COVID-19. Cytokine storms, which can rapidly cause single or multiple organ failure and ultimately can be life-threatening, are considered to be an important cause of death in patients with severe COVID-19” [2]. Another research study showed that “this cohort of hospitalized patients with COVID-19 in a minority-predominant population, severe obesity (BMI ≥ 35 kg/m2), increasing age, and male sex were independently associated with higher in-hospital mortality and in general worse in-hospital outcomes” [3]. Some researchers also proved “obesity is associated with a nearly 3-fold increased risk for severe COVID-19 with a dose-effect relationship between increasing BMI and the proportion of patients with severe illness. In the current analysis, the risk of severe illness in MAFLD patients with co-existing obesity was >6-fold greater after adjustment for confounders” [6]. This also indicated that obesity as well as increasing BMI are key indicators/features on the severity of COVID-19 [6]. In the treatment for COVID-19, it was pointed out that “patients who received invasive mechanical ventilation were more likely to be male, to have obesity, and to have elevated liver-function values and inflammatory markers (ferritin, d-dimer, C-reactive protein, and procalcitonin) than were patients who did not receive invasive mechanical ventilation” [4].

In a clinical study conducted in Shanghai, China, it suggested a list of foods and their impacts on our white blood cell and C-reactive protein (CRP) levels, which in term could affect our health situations [5]. It listed that “red meat, refined carbohydrates like white bread and pasta, fried foods, and products with added sugars, such as sugary drinks” should be avoided if your CRP is high [5]. With that being said, it also recommended “on-pharmacological methods of reducing CRP include aerobic exercise, smoking cessation, weight loss, and a heart-healthy diet. In other words, taking aggressive steps to make your lifestyle healthier will also result in a reduced CRP level” [5].

**III. METHODOLOGY**

Dataset:

We used the COVID-19 Healthy Diet Dataset. This dataset (https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset) has combined data of different food types, world population obesity and undernourished rate, and global COVID-19 cases count from numerous countries. Included in the dataset are fat quantity, energy intake (kcal), food supply quantity (kg), and protein for different categories of food (all calculated as percentage of total intake amount). Obesity and undernourished rate (also in percentage) have also been added for comparison.

Preprocessing:

We removed the last column, population, as well as all incomplete cases from other columns. We assessed the correlations between all numerical columns to find strong positive and negative relationships, ranging from -1 to 1. We zoned in on the correlations between obesity, deaths, confirmed and recovered. Using multiple linear regressions, we fit obesity with all the other variables to predict the value of the outcome variable Y (Obesity) based on the input predictor variables X (All the other variables). The model is statistically significant when the p-Values are less than 0.05 (95% Confidence). But if P >|t|, that means the coefficients are not significant. We fit the linear regression model to be used in a stepwise algorithm. We were able to choose a model by AIC in a Stepwise Algorithm, which is a way to build a model by adding or removing predictor variables. For KNN preprocessing, we dropped the last column which is the unit (all except population) because it had just a percentage sign which was no use to us. Then we checked for null values by summing the null values. Obesity, undernourished, confirmed, deaths, recovered, active had null values so we dropped them. The column undernourished was labeled as an object data type (categorical) so we needed to convert it to float64 (Numeric). The undernourished Column has values of <2.5 so I replaced the <2.5 with 2.0 changing it to numeric.

Classifier Design:

We chose Logistic Regression, K-Nearest Neighbor and Random Forest as our supervised learning classifiers. We trained our models and then evaluated them on the same dataset with our test dataset. We split the data into 70% train 30% test.

**IV. EXPERIMENTS:**

Logistic Regression:

Two sets of logistic regression models have been built and trained with several important features. One built with the feature Obesity and three other dependent variables, Deaths, Recovered, Confirmed was to show the correlations of obesity among the end results of COVID patients, and its evaluation with Mean Squared Error (MSE) for Recovered~Obesity was 0.044 being the lowest. Another one built with the features Sugar (including variables Sugar.Crops and Sugar...Sweeteners), Animal Fats, Vegetable Oils and the dependent variable Obesity was to show the correlations of obesity among three different nutritional intakes to seek how diets are correlated with obesity, and its evaluation with MSE for Obesity~Vegetable Oil was 17.91 being the highest.

The p-values of two sets of models would indicate whether the results are less likely to occur by chance. Through the Residuals vs. Fitted plot, based on the curve pattern, we could indicate whether there are non-linear relationships. The Normal Q-Q plot will help us indicate whether the residuals are normally distributed. The Scale-Location plot will assist us to figure out whether the residuals have equal variance along the regression line, based on how the data is spread (above or below) in the space of the regression line. The Residuals vs. Leverage plot was used to find out cases that have high leverage, or influence, in the model.

## Decision Tree:

## We used obesity to construct the tree (Recovered~ Obesity) and used the ANOVA method to gain information about the relationship between the dependent (Recovered) and independent variable (Obesity). We printed the complexity parameter (CP) table that displays the decision tree results and a CP plot allows us to visualize the cross-validation results of optimal prunings. CP is used to control and select the optimal tree size, a total of 6 nodes were created. The x-error is the cross-validation error, you use the 3 columns: rel\_error, xerror and xstd to help choose where to prune the tree and each row represents a different height of the tree.

Evaluation with Mean Squared Error For Logistic Regression and Decision Tree:

The MSE is a measure of the quality of an estimator; values closer to zero are better. Mean squared error is the difference between the estimated values and the actual results. Our expected results differ from this prediction which is a slight error from the estimated value. For the decision tree, MSE is only 1.75 (Recovered~ Obesity), 0 means the model has perfect accuracy, so our decision tree model is a good fit. For one set of logistic regressions, MSE is only 0.044 for (Obesity~ Deaths), so our logistic regression model between obesity and death is a good fit. For the second set of logistic regressions, the lowest MSE is 17.91 for the Obesity~Vegetable Oil model, so our logistic regression model between obesity and vegetable oil is not a good fit.

KNN:

We can do a couple of modeling exercises with this dataset. The first thing we’ll try is to check if we can predict the obesity above average column which we added using food features. We draw a heat map from the co relations and we see there is a vertical line so there is a strong correlation between those elements.

Now for features we will not include the deaths because that is our target. Our model features include alcoholic beverages, animal products, animal fats, milk, obesity etc. We do some target balancing to convert the bad diet food features into float (numeric values). X will = features and Y will = recovered. We import train test split from sklearn.model\_selection

After we do some data preprocessing on the training set. As there are no missing values, we will build a pipeline to scale features to have similar maginutes by bringing all of them between 0-1 using MinMaxScaler and then apply the KNN classifier. I use K=3 as the estimator. Then we train our classifier with the .fit() method. Using the fitted model we make a prediction on the training set giving us a high train accuracy of 97,6%. Then we draw the confusion matrix of that. Now we check the model on the test set which leads us to an accuracy of 87.1%. We also draw the confusion matrix which shows the true positives (18) , true negatives (9) , false positives (3) and false negatives (1). From this classification method, we can conclude that there are similarities between obesity, bad diet food features, and number of recovered and deaths.

|  |  |
| --- | --- |
| **Model (Tuned)** | **Test Accuracy** |
| Logistic Regression | 0.044 MSE |
| Decision Tree | 1.75 MSE |
| KNN classifier | 87% |

**V. CONCLUSION:**

This analysis still proves evidence that a healthier diet is an important step that we can take to help reduce the amount of COVID cases and deaths.We conclude that there are similarities between obesity, bad diet food features, and number of recovered and deaths. In order to stay healthy, we need to eat healthy. Hopefully, the results will help countless Coronavirus-confirmed patients, researchers, and doctors around the world to overcome them together.

**Reference**

*[1] Ebinger, Joseph E., et al. “Pre-Existing Traits Associated with Covid-19 Illness Severity.” MedRxiv, Cold Spring Harbor Laboratory Press, 1 Jan. 2020, www.medrxiv.org/ content/10.1101/2020.04.29.20084533v2.*

*[2] Liu, Tao, et al. “The Potential Role of IL-6 in Monitoring Severe Case of Coronavirus Disease 2019.” MedRxiv, Cold Spring Harbor Laboratory Press, 1 Jan. 2020,*

*www.medrxiv.org/content/10.1101/2020.03.01.20029769v2.full-text.*

*[3] Palaiodimos, Leonidas et al. “Severe obesity, increasing age and male sex are*

*independently associated with worse in-hospital outcomes, and higher in-hospital*

*mortality, in a cohort of patients with COVID-19 in the Bronx, New York.” Metabolism: clinical and experimental vol. 108 (2020): 154262.*

*doi:10.1016/j.metabol.2020.154262*

*[4] T. T. Shimabukuro and Others, et al. “Clinical Characteristics of Covid-19 in New York City: NEJM.” New England Journal of Medicine, 21 Apr. 2021, www.nejm.org/doi/10. 1056/NEJMc2010419.*

*[5] Yu X, Sun X, Cui P, Pan H, Lin S, Han R, Jiang C, Fang Q, Kong D, Zhu Y, Zheng Y, Gong X, Xiao W, Mao S, Jin B, Wu H, Fu C. Epidemiological and clinical characteristics of 333 confirmed cases with coronavirus disease 2019 in Shanghai, China. Transbound*

*Emerg Dis. 2020 Jul;67(4):1697-1707. doi: 10.1111/tbed.13604. Epub 2020 May 13. PMID: 32351037; PMCID: PMC7267440.*

*[6] Zheng, Kenneth I., et al. “Letter to the Editor: Obesity as a Risk Factor for Greater*

*Severity of COVID-19 in Patients with Metabolic Associated Fatty Liver Disease.”*

*Metabolism, vol. 108, 2020, p. 154244., doi:10.1016/j.metabol.2020.154244.*