COVID ML Program

Machine Learning Algorithm:

The goal of this algorithm is to examine patient conditions and assign a final outcome to each patient. By feeding the algorithm significant data, we can extrapolate how likely it is a patient will die from COVID-19 or will require intensive care. By presenting this data accurately, it will serve as a model to aid doctors in their fight against the pandemic.

Data Analysis

COVID-19 is an insidious disease. It can pry on the most vulnerable populations as well as cause a considerable impact on healthy individuals.

Based on current research the following conditions have been shown to have a significant impact on COVID Mortality. These will make up the basis of the program.

Dataset Summary:

Age

Blood Pressure - Systolic only

Smoker - Y/N

O2 Saturation - %

Asthma - Y/N

Diabetes - Y/N

Obesity - Y/N

Lung Disease - Y/N

Heart Disease - Y/N

Program Execution

How the program will execute its functions:

By applying a k-th nearest neighbor algorithm, I will be able to give within reasonable accuracy the chance of mortality. This result will be derived from the Machine Learning Program as it will be able to assess the greatest risk factors and associate patient data with a higher likelihood of mortality.

To ensure accuracy, I will also be implementing a secondary ML program to verify/contradict the results. Any discrepancies will be obvious, and will allow for accurate assessment of reasoning behind the results.

Data Creation

Given the broad range of COVID symptoms and the resulting effects (i.e 02 Saturation, BP, and disease history), it is important to normalize the data. This will ensure that outliers are included in the final analysis. It also establishes a baseline for the algorithm to work off of to ensure a similar picture and accurate data representation.

Research has shown that systolic blood pressure is a more accurate predictor of risk. A problem noticed in creation was the varying range blood pressure where the ratio of a healthy person could be higher than that of an unhealthy person, where it should be the opposite. This leads to too much variance and no specific trend being noticed.

```
library(class)
#PROBABILITY DETERMINED AS MAJORITY OF THE POPULATION FALLS JUST OUT OF HEALTHY RANGE- PROB DECREASES F
BP<-sample(c(120, 125, 130, 143, 155), size = 100, replace = TRUE,prob = c(.3, .4, .2, .1, .1))
Smoker<-sample(c(1,0), size = 100, replace = TRUE, prob = c(.15, .85)) #Current smoker population is 15
Age<-sample(x = 19:80, size = 100, replace = TRUE)
O2_sat <-sample(x = 89:99, size = 100, replace = TRUE) #For simplicity - Equal probability given, lack
Asthma <-sample(c(1,0), size = 100, replace= TRUE, prob = c(.07, .93)) #Reflective of Population
Mortality <- sample(c(1,0), size = 100, replace = TRUE, prob = c(.3, .7)) #For efficiency, and proof of
FINAL_DF <- data.frame(BP, Smoker, Age,O2_sat, Asthma, Mortality)
```

```
out <- knn(KNN_train[,1:6],
           KNN_test[,1:6],
           cl = KNN_train$Mortality,
           k=20)

LearnTable <- table(out, KNN_test$Mortality)
LearnTable
```

```
##
## out                 -0.62792174216674 1.56980435541685
##   -0.62792174216674                21                8
##   1.56980435541685                  0                0
```

```
accuracy<-function(x){sum(diag(x)/sum(rowSums(x)))*100}
accuracy(LearnTable)
```

```
## [1] 72.41379
```

Accuracy will vary due to the randomization. Due to the binary nature Mortality- Column 1 and 2 are mislabeled- 1st Column is Survival, 2nd is Death.

Probit Regression Model

A K-Model serves to group similar variables together based on specified parameter. However, in our case we hope to accurately predict the outcome of a patient. Accurately modeling this requires a ML program that assesses the outcome (dependent) and predictor variables(independent).

```
probit <- glm(formula = Mortality ~ BP + Smoker + Age + O2_sat + Asthma, family= binomial(link = "probi

summary(probit)
```

```
##
## Call:
## glm(formula = Mortality ~ BP + Smoker + Age + O2_sat + Asthma,
##     family = binomial(link = "probit"), data = FINAL_DF)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.333  -0.771  -0.696   1.172   1.843
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.4263172  4.5344396  -0.756    0.450
## BP           0.0213956  0.0138504   1.545    0.122
## Smoker       0.2638071  0.3526188   0.748    0.454
## Age          0.0009722  0.0078848   0.123    0.902
## O2_sat      -0.0008090  0.0429300  -0.019    0.985
## Asthma       0.7589942  0.4750751   1.598    0.110
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118.59  on 99  degrees of freedom
## Residual deviance: 113.90  on 94  degrees of freedom
## AIC: 125.9
##
## Number of Fisher Scoring iterations: 4
```

Due to the nature of our data set, we can clearly see that these variables do not accurately reflect that of the current pandemic. A data set with accurate results (i.e smokers have a higher chance of Mortality, as well as: Asthma, high BP, and Age) from real data sets would give accurate readings. But as this is a proof of concept-we will ignore.

```
xtabs(~Mortality + BP + Smoker, data = FINAL_DF)
```

```
## , , Smoker = 0
##
##          BP
## Mortality 120 125 130 143 155
##         0  20  22  10   3   5
##         1   3   8   7   2   2
##
## , , Smoker = 1
##
##          BP
## Mortality 120 125 130 143 155
##         0   1   9   2   0   0
##         1   2   2   1   0   1
```

This short summary allows us to determine trends in smokers vs non smokers, as well as their BP and the corresponding mortality rates.

Final Summary

To improve on these models, a significant amount of changes need to be made: Larger data Set (>10000), accurate reflection of population percentages, and defined mortality rates.

While a work in progress, I believe this demonstrates great utility.