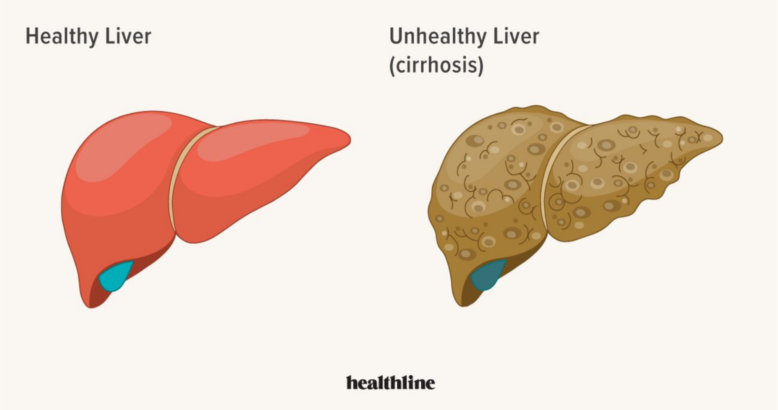
Liver disease predictions



Hanisha Pasupulate, Robert Jenkins, Nick Killian, Sheikh-Sedat Touray, Andrew Bessire

DSP 556

Data overview

- Data gathered from South Korea spanning 2002-2021 through the South Korean National Health Insurance Service

- An annual random sampling of ~ 1 million individuals who underwent basic yearly check-ups covered by the South Korean National Health Insurance

- Dataset acquired via Kaggle

LITERATURE

- 4.5 million adults in the US are diagnosed with Liver disease each year
- Link between lipid profile (LDL, HDL, triglycerides) and liver disease observed in patients with cirrhosis
- Serum lipid levels serve as crucial indicators of liver damage
- GGT (glutamyltransferase) is a powerful predictor of liver disease
- Alcohol consumption and obesity as significant risk factors for developing cirrhosis

2009 vs 2019

- Analyzed data from 2009 and 2019

- The selection of 2009 and 2019 was deliberate to ensure a significant gap between the two sets of collected data

- Additionally, the exclusion of 2020 was intentional due to the disruptive impact of the COVID-19 pandemic.

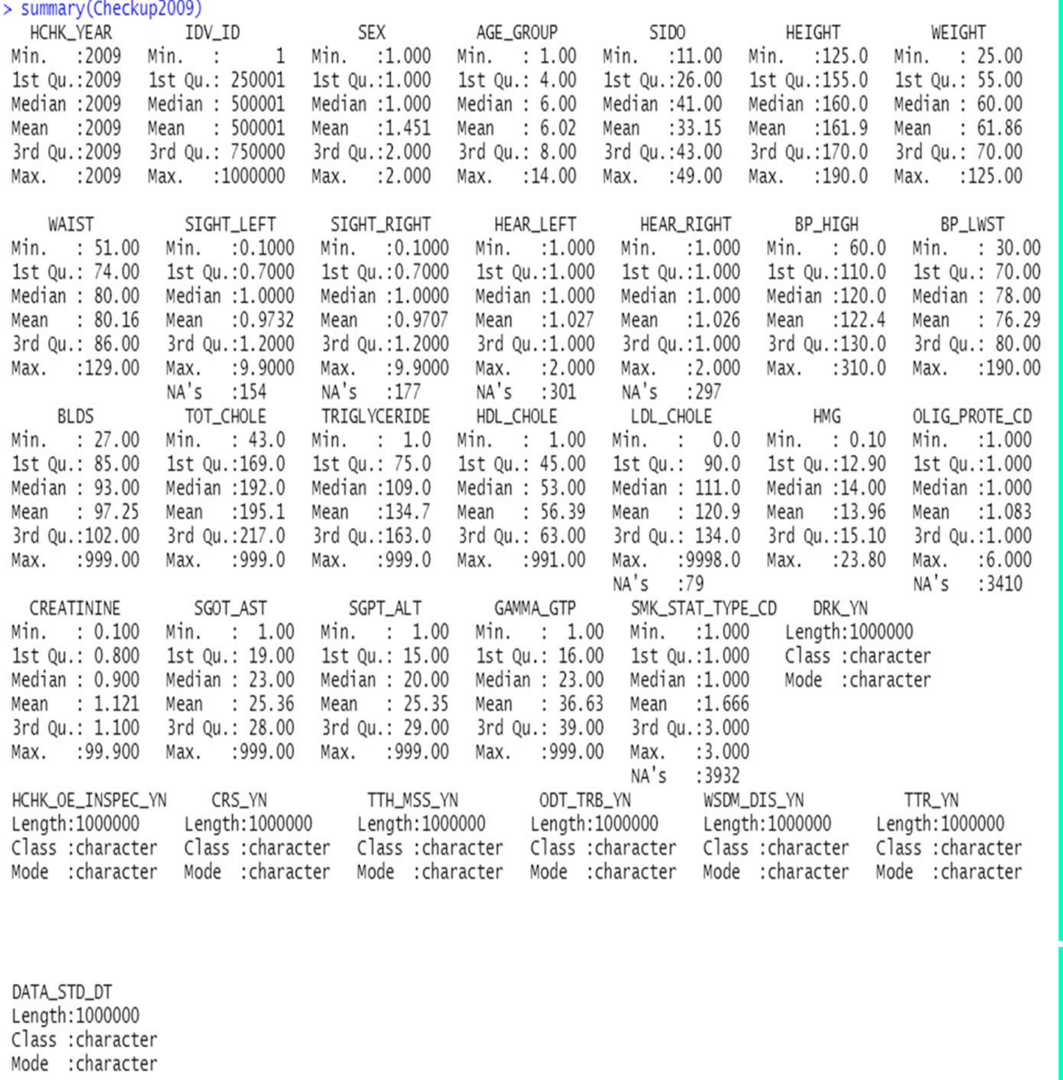
DATA

- Variable Explanations
  - Sex - M/F
  - Age - groups 1-18 for every 5 years
  - Height - centimeters
  - Weight - kilograms
  - BP_High & BP_LWST - Systolic and Diastolic Blood Pressure (N/mm Hg)
  - BLDS - Blood Glucose level per 100ml of blood (N/mg/dL)
  - TOT_CHOLE - Sum of ester and non-ester cholesterol (mg/dL)
    - Normal values 150-250 mg/dL
  - HDL_CHOLE - Cholesterol level in HDL (mg/dL
    - Normal values 30-65 mg/dL
  - Triglyceride - amount of simple or neutral lipids (mg/dL)
    - Normal values 30-135 mg/dL
  - HMG - level of pigment protein in blood, helps carry oxygen (g/dL)
  - SGOT_AST - blood test level that indicates liver function (IU/L)
    - Normal values 0-40 IU/L, high concentration can indicate organ damage

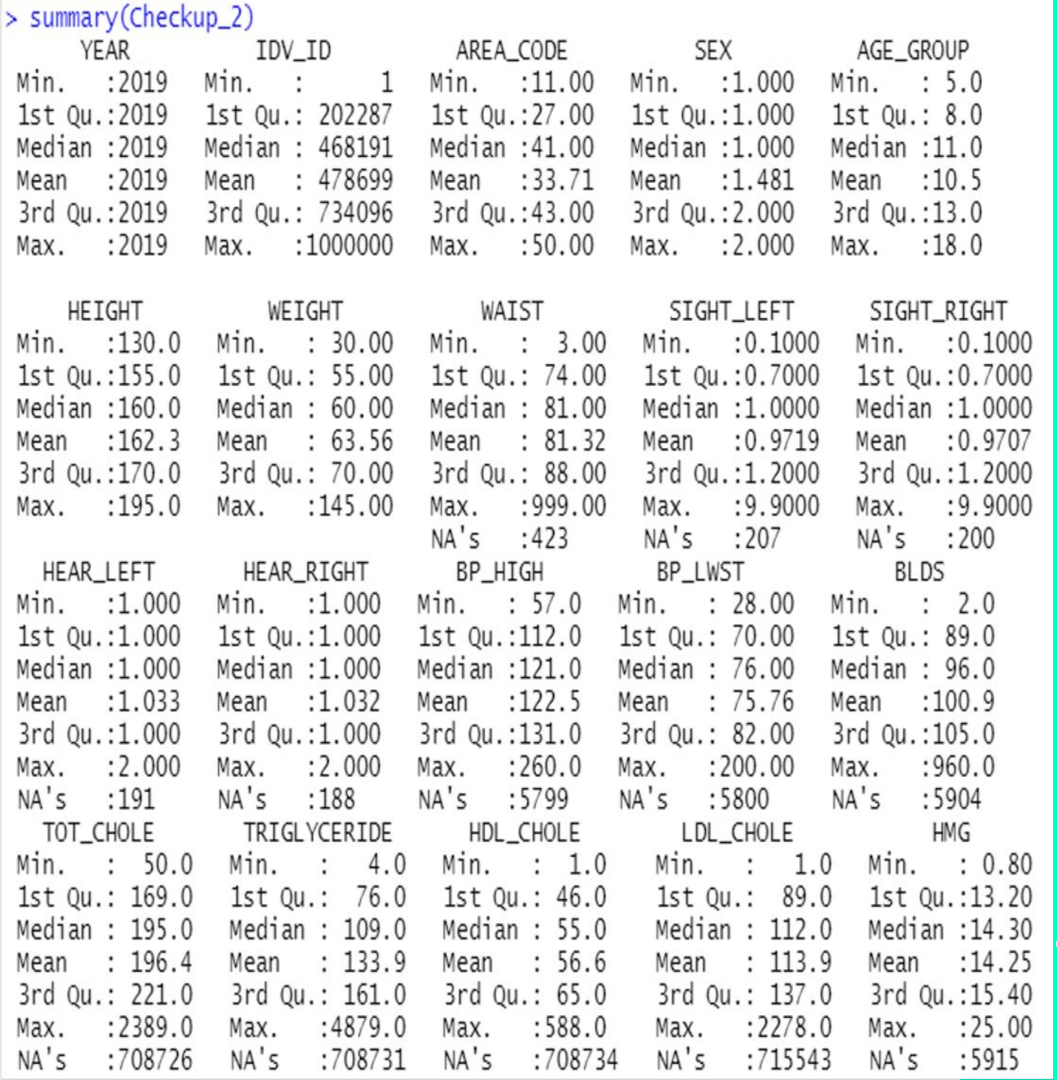
Goal

Generate machine learning models that utilize different health indicators to predict a patient’s liver enzyme levels and thus their chance of possible liver disease

2009 Summary stats



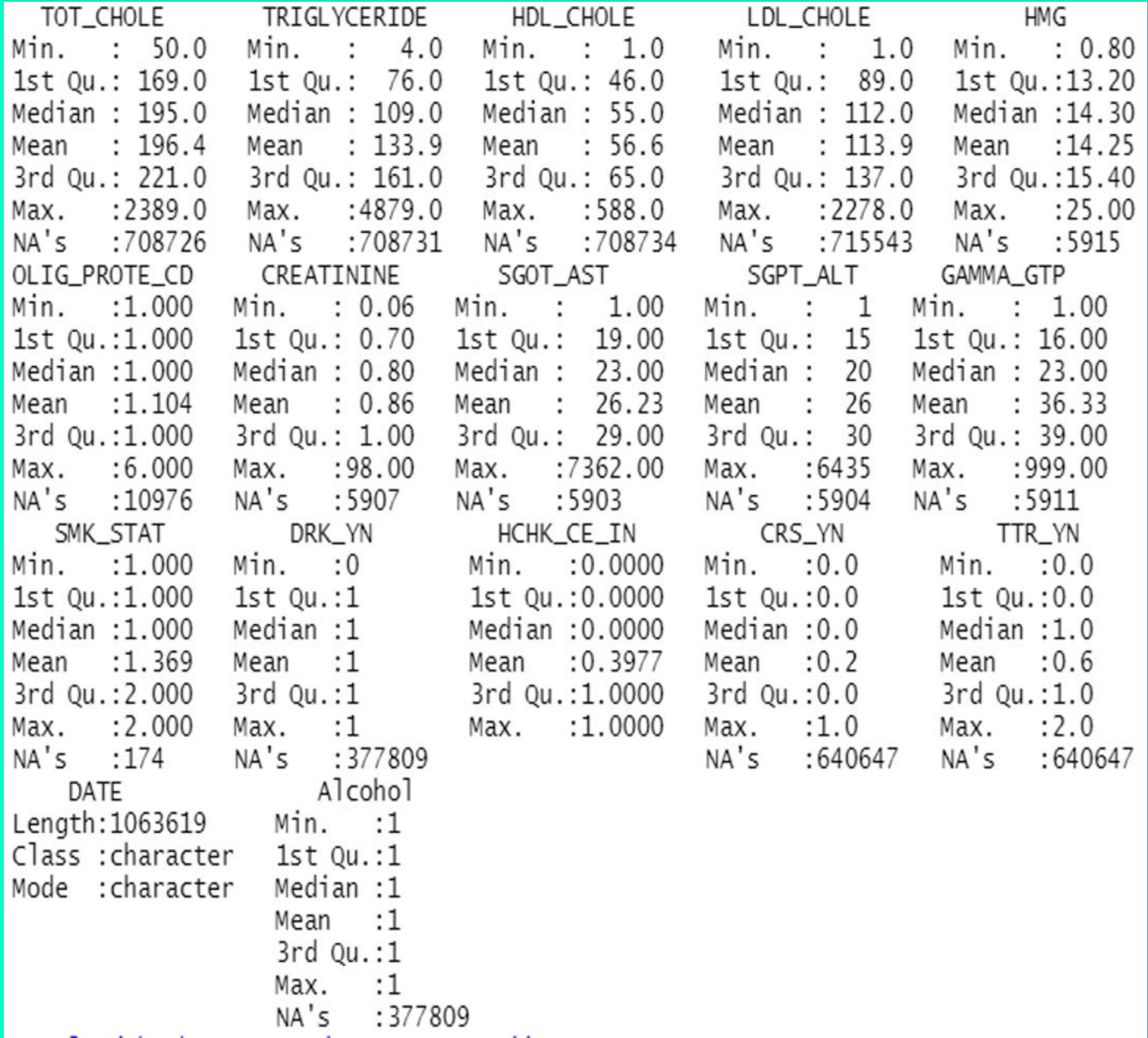
```
> summary(Checkup2009)
   HCHK_YEAR       IDV_ID            SEX          AGE_GROUP         SIDO           HEIGHT          WEIGHT     
 Min.   :2009   Min.   :      1   Min.   :1.000   Min.   : 1.00   Min.   :11.00   Min.   :125.0   Min.   : 25.00  
 1st Qu.:2009   1st Qu.: 250001   1st Qu.:1.000   1st Qu.: 4.00   1st Qu.:26.00   1st Qu.:155.0   1st Qu.: 55.00  
 Median :2009   Median : 500001   Median :1.000   Median : 6.00   Median :41.00   Median :160.0   Median : 60.00  
 Mean   :2009   Mean   : 500001   Mean   :1.451   Mean   : 6.02   Mean   :33.15   Mean   :161.9   Mean   : 61.86  
 3rd Qu.:2009   3rd Qu.: 750000   3rd Qu.:2.000   3rd Qu.: 8.00   3rd Qu.:43.00   3rd Qu.:170.0   3rd Qu.: 70.00  
 Max.   :2009   Max.   :1000000   Max.   :2.000   Max.   :14.00   Max.   :49.00   Max.   :190.0   Max.   :125.00  
                                                                                                                  
     WAIST          SIGHT_LEFT       SIGHT_RIGHT       HEAR_LEFT       HEAR_RIGHT        BP_HIGH          BP_LWST      
 Min.   : 51.00   Min.   :0.1000   Min.   :0.1000   Min.   :1.000   Min.   :1.000   Min.   : 60.0   Min.   : 30.00  
 1st Qu.: 74.00   1st Qu.:0.7000   1st Qu.:0.7000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:110.0   1st Qu.: 70.00  
 Median : 80.00   Median :1.0000   Median :1.0000   Median :1.000   Median :1.000   Median :120.0   Median : 78.00  
 Mean   : 80.16   Mean   :0.9732   Mean   :0.9707   Mean   :1.027   Mean   :1.026   Mean   :122.4   Mean   : 76.29  
 3rd Qu.: 86.00   3rd Qu.:1.2000   3rd Qu.:1.2000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:130.0   3rd Qu.: 80.00  
 Max.   :129.00   Max.   :9.9000   Max.   :9.9000   Max.   :2.000   Max.   :2.000   Max.   :310.0   Max.   :190.00  
                  NA's   :154      NA's   :177      NA's   :301     NA's   :297                                      
      BLDS           TOT_CHOLE      TRIGLYCERIDE      HDL_CHOLE        LDL_CHOLE           HMG          OLIG_PROTE_CD  
 Min.   : 27.00   Min.   : 43.0   Min.   :  1.0   Min.   :  1.00   Min.   :   0.0   Min.   : 0.10   Min.   :1.000  
 1st Qu.: 85.00   1st Qu.:169.0   1st Qu.: 75.0   1st Qu.: 45.00   1st Qu.:  90.0   1st Qu.:12.90   1st Qu.:1.000  
 Median : 93.00   Median :192.0   Median :109.0   Median : 53.00   Median : 111.0   Median :14.00   Median :1.000  
 Mean   : 97.25   Mean   :195.1   Mean   :134.7   Mean   : 56.39   Mean   : 120.9   Mean   :13.96   Mean   :1.083  
 3rd Qu.:102.00   3rd Qu.:217.0   3rd Qu.:163.0   3rd Qu.: 63.00   3rd Qu.: 134.0   3rd Qu.:15.10   3rd Qu.:1.000  
 Max.   :999.00   Max.   :999.0   Max.   :999.0   Max.   :991.00   Max.   :9998.0   Max.   :23.80   Max.   :6.000  
                                                                    NA's   :79                        NA's   :3410   
   CREATININE        SGOT_AST         SGPT_ALT         GAMMA_GTP      SMK_STAT_TYPE_CD     DRK_YN         
 Min.   : 0.100   Min.   :  1.00   Min.   :  1.00   Min.   :  1.00   Min.   :1.000    Length:1000000    
 1st Qu.: 0.800   1st Qu.: 19.00   1st Qu.: 15.00   1st Qu.: 16.00   1st Qu.:1.000    Class :character  
 Median : 0.900   Median : 23.00   Median : 20.00   Median : 23.00   Median :1.000    Mode  :character  
 Mean   : 1.121   Mean   : 25.36   Mean   : 25.35   Mean   : 36.63   Mean   :1.666                      
 3rd Qu.: 1.100   3rd Qu.: 28.00   3rd Qu.: 29.00   3rd Qu.: 39.00   3rd Qu.:3.000                      
 Max.   :99.900   Max.   :999.00   Max.   :999.00   Max.   :999.00   Max.   :3.000                      
                                                                     NA's   :3932                       
 HCHK_OE_INSPEC_YN     CRS_YN           TTH_MSS_YN         ODT_TRB_YN         WSDM_DIS_YN          TTR_YN         
 Length:1000000     Length:1000000     Length:1000000     Length:1000000     Length:1000000     Length:1000000    
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                                                                  
                                                                                                                  
                                                                                                                  
                                                                                                                  
 DATA_STD_DT       
 Length:1000000    
 Class :character  
 Mode  :character  
```



```
> summary(Checkup_2)
      YEAR          IDV_ID          AREA_CODE          SEX          AGE_GROUP   
 Min.   :2019   Min.   :      1   Min.   :11.00   Min.   :1.000   Min.   : 5.0  
 1st Qu.:2019   1st Qu.: 202287   1st Qu.:27.00   1st Qu.:1.000   1st Qu.: 8.0  
 Median :2019   Median : 468191   Median :41.00   Median :1.000   Median :11.0  
 Mean   :2019   Mean   : 478699   Mean   :33.71   Mean   :1.481   Mean   :10.5  
 3rd Qu.:2019   3rd Qu.: 734096   3rd Qu.:43.00   3rd Qu.:2.000   3rd Qu.:13.0  
 Max.   :2019   Max.   :1000000   Max.   :50.00   Max.   :2.000   Max.   :18.0  

     HEIGHT          WEIGHT           WAIST          SIGHT_LEFT       SIGHT_RIGHT    
 Min.   :130.0   Min.   : 30.00   Min.   :  3.00   Min.   :0.1000   Min.   :0.1000  
 1st Qu.:155.0   1st Qu.: 55.00   1st Qu.: 74.00   1st Qu.:0.7000   1st Qu.:0.7000  
 Median :160.0   Median : 60.00   Median : 81.00   Median :1.0000   Median :1.0000  
 Mean   :162.3   Mean   : 63.56   Mean   : 81.32   Mean   :0.9719   Mean   :0.9707  
 3rd Qu.:170.0   3rd Qu.: 70.00   3rd Qu.: 88.00   3rd Qu.:1.2000   3rd Qu.:1.2000  
 Max.   :195.0   Max.   :145.00   Max.   :999.00   Max.   :9.9000   Max.   :9.9000  
                                  NA's   :423      NA's   :207      NA's   :200     
   HEAR_LEFT       HEAR_RIGHT       BP_HIGH         BP_LWST            BLDS      
 Min.   :1.000   Min.   :1.000   Min.   : 57.0   Min.   : 28.00   Min.   :  2.0  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:112.0   1st Qu.: 70.00   1st Qu.: 89.0  
 Median :1.000   Median :1.000   Median :121.0   Median : 76.00   Median : 96.0  
 Mean   :1.033   Mean   :1.032   Mean   :122.5   Mean   : 75.76   Mean   :100.9  
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:131.0   3rd Qu.: 82.00   3rd Qu.:105.0  
 Max.   :2.000   Max.   :2.000   Max.   :260.0   Max.   :200.00   Max.   :960.0  
 NA's   :191     NA's   :188     NA's   :5799    NA's   :5800     NA's   :5904   
   TOT_CHOLE       TRIGLYCERIDE      HDL_CHOLE         LDL_CHOLE           HMG       
 Min.   :  50.0   Min.   :   4.0   Min.   :  1.0   Min.   :   1.0   Min.   : 0.80  
 1st Qu.: 169.0   1st Qu.:  76.0   1st Qu.: 46.0   1st Qu.:  89.0   1st Qu.:13.20  
 Median : 195.0   Median : 109.0   Median : 55.0   Median : 112.0   Median :14.30  
 Mean   : 196.4   Mean   : 133.9   Mean   : 56.6   Mean   : 113.9   Mean   :14.25  
 3rd Qu.: 221.0   3rd Qu.: 161.0   3rd Qu.: 65.0   3rd Qu.: 137.0   3rd Qu.:15.40  
 Max.   :2389.0   Max.   :4879.0   Max.   :588.0   Max.   :2278.0   Max.   :25.00  
 NA's   :708726   NA's   :708731   NA's   :708734   NA's   :715543   NA's   :5915   
```

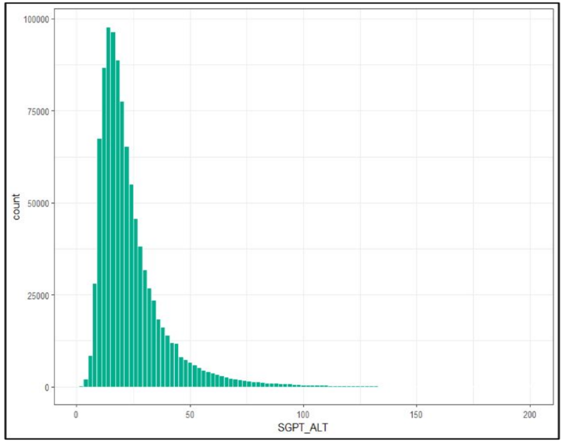
2019
Summary stats

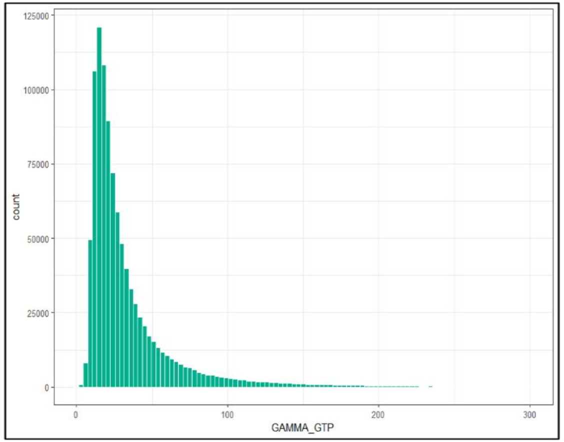
2019 Summary stats

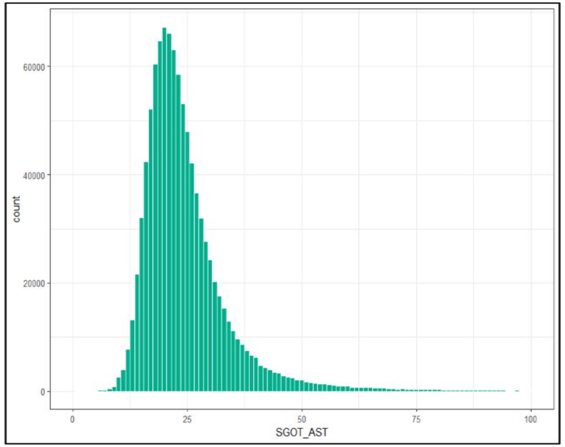


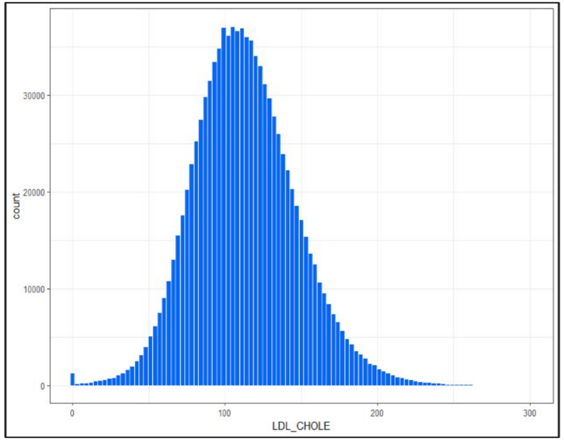
```
  TOT_CHOLE       TRIGLYCERIDE      HDL_CHOLE       LDL_CHOLE          HMG       
Min.   :  50.0   Min.   :   4.0   Min.   :  1.0   Min.   :   1.0   Min.   : 0.80  
1st Qu.: 169.0   1st Qu.:  76.0   1st Qu.: 46.0   1st Qu.:  89.0   1st Qu.:13.20  
Median : 195.0   Median : 109.0   Median : 55.0   Median : 112.0   Median :14.30  
Mean   : 196.4   Mean   : 133.9   Mean   : 56.6   Mean   : 113.9   Mean   :14.25  
3rd Qu.: 221.0   3rd Qu.: 161.0   3rd Qu.: 65.0   3rd Qu.: 137.0   3rd Qu.:15.40  
Max.   :2389.0   Max.   :4879.0   Max.   :588.0   Max.   :2278.0   Max.   :25.00  
NA's   :708726   NA's   :708731   NA's   :708734  NA's   :715543   NA's   :5915   
OLIG_PROTE_CD      CREATININE       SGOT_AST          SGPT_ALT       GAMMA_GTP     
Min.   :1.000   Min.   : 0.06   Min.   :   1.00   Min.   :   1   Min.   :  1.00  
1st Qu.:1.000   1st Qu.: 0.70   1st Qu.:  19.00   1st Qu.:  15   1st Qu.: 16.00  
Median :1.000   Median : 0.80   Median :  23.00   Median :  20   Median : 23.00  
Mean   :1.104   Mean   : 0.86   Mean   :  26.23   Mean   :  26   Mean   : 36.33  
3rd Qu.:1.000   3rd Qu.: 1.00   3rd Qu.:  29.00   3rd Qu.:  30   3rd Qu.: 39.00  
Max.   :6.000   Max.   :98.00   Max.   :7362.00   Max.   :6435   Max.   :999.00  
NA's   :10976   NA's   :5907    NA's   :5903      NA's   :5904   NA's   :5911    
   SMK_STAT         DRK_YN        HCHK_CE_IN         CRS_YN          TTR_YN     
Min.   :1.000   Min.   :0      Min.   :0.0000   Min.   :0.0     Min.   :0.0     
1st Qu.:1.000   1st Qu.:1      1st Qu.:0.0000   1st Qu.:0.0     1st Qu.:0.0     
Median :1.000   Median :1      Median :0.0000   Median :0.0     Median :1.0     
Mean   :1.369   Mean   :1      Mean   :0.3977   Mean   :0.2     Mean   :0.6     
3rd Qu.:2.000   3rd Qu.:1      3rd Qu.:1.0000   3rd Qu.:0.0     3rd Qu.:1.0     
Max.   :2.000   Max.   :1      Max.   :1.0000   Max.   :1.0     Max.   :2.0     
NA's   :174     NA's   :377809                  NA's   :640647  NA's   :640647  
    DATE               Alcohol      
Length:1063619     Min.   :1        
Class :character   1st Qu.:1        
Mode  :character   Median :1        
                   Mean   :1        
                   3rd Qu.:1        
                   Max.   :1        
                   NA's   :377809   
```

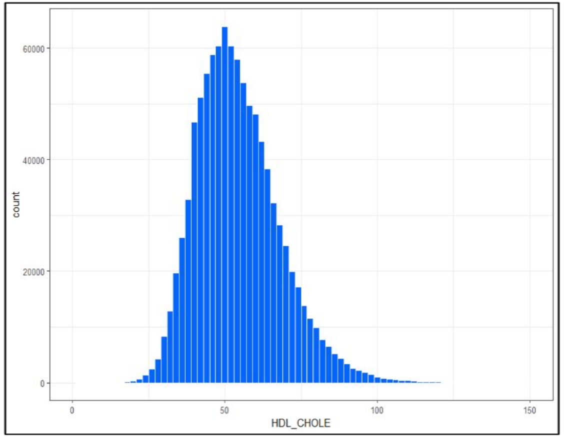
2009 Key Variables

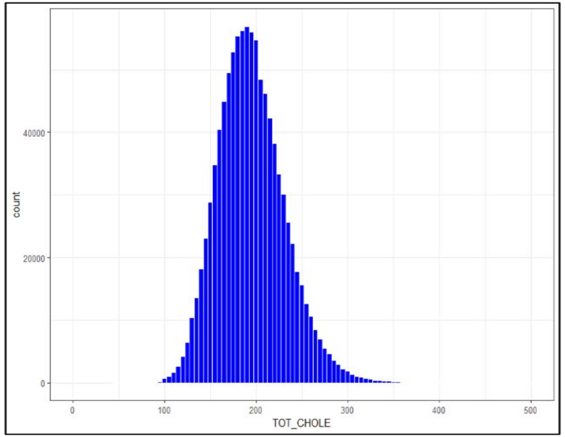




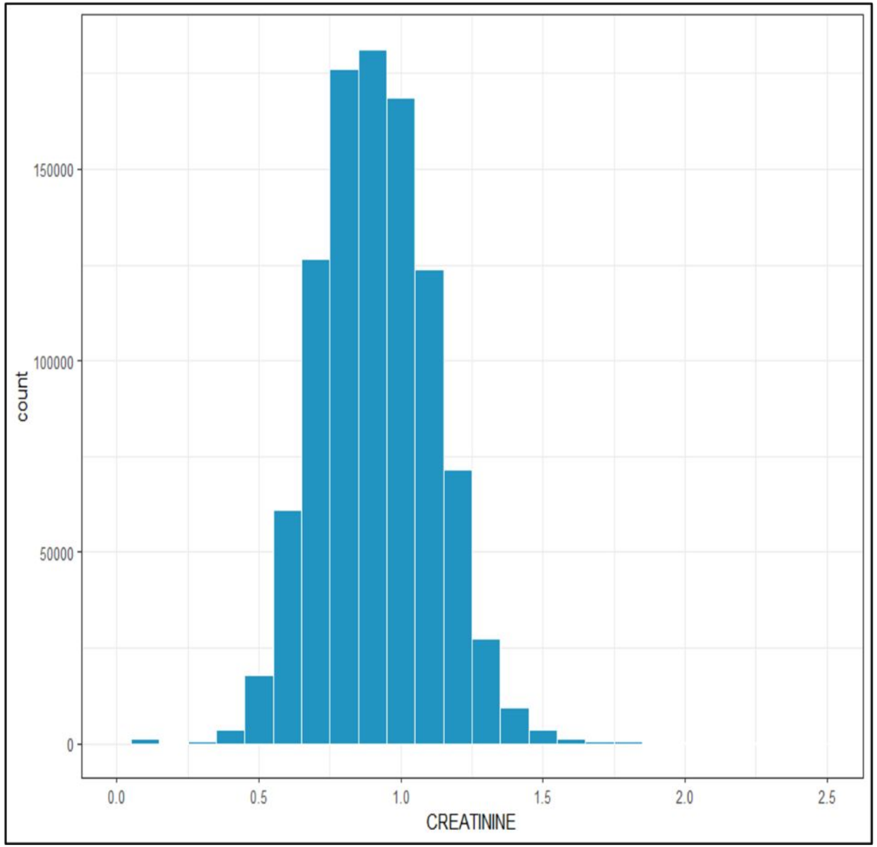


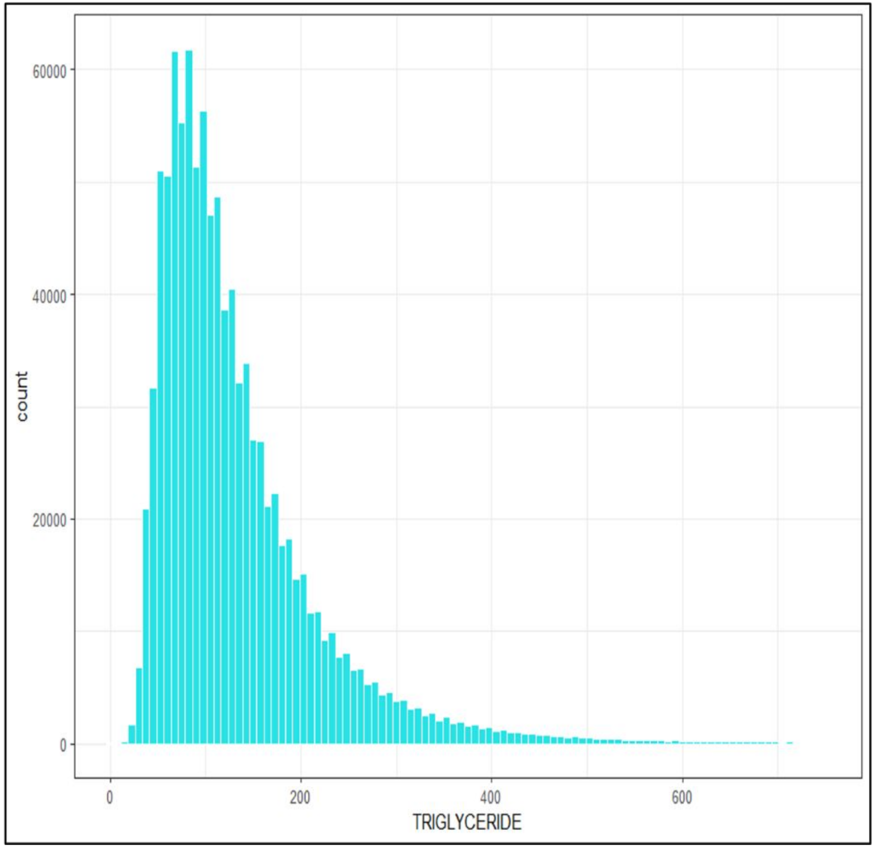




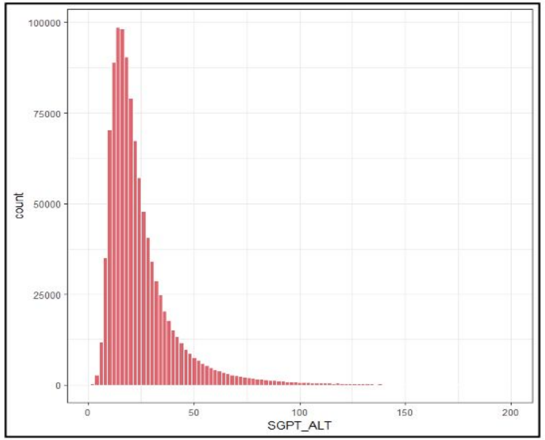


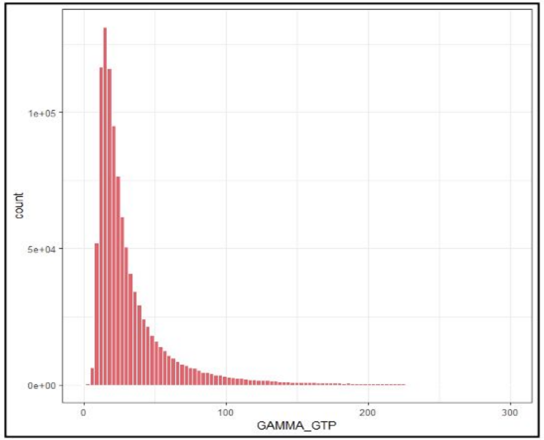
2009 Key variables

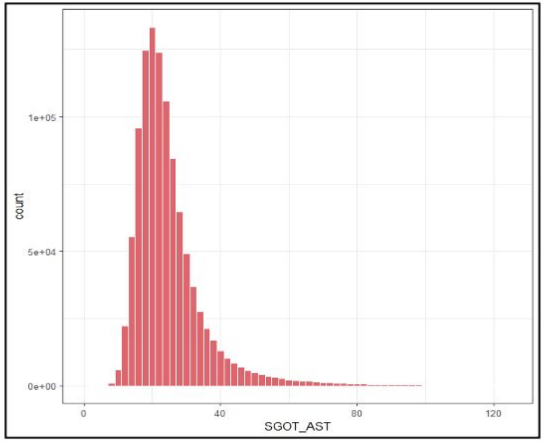


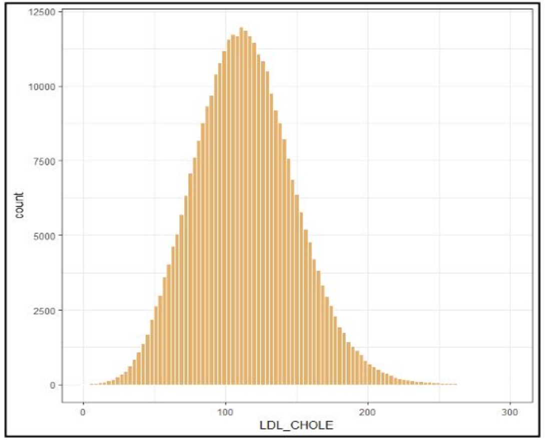


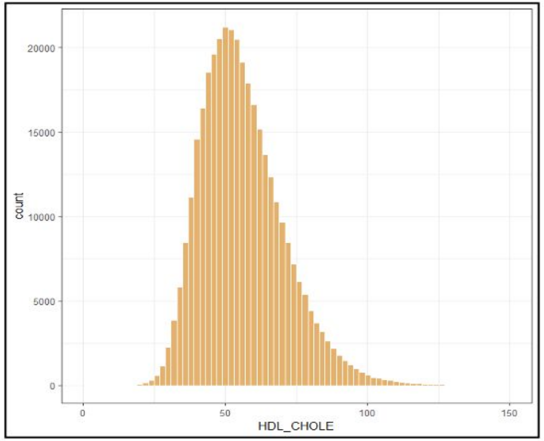
2019 Key Variables

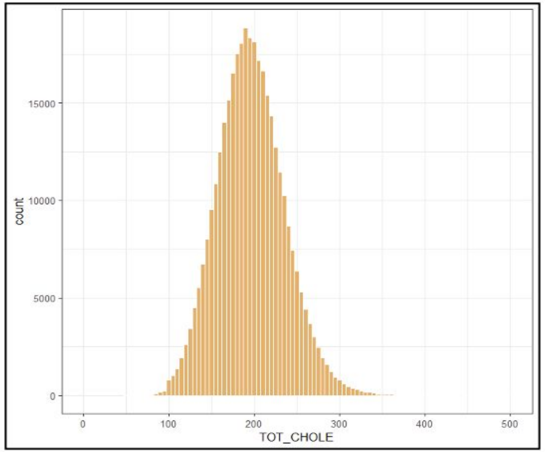




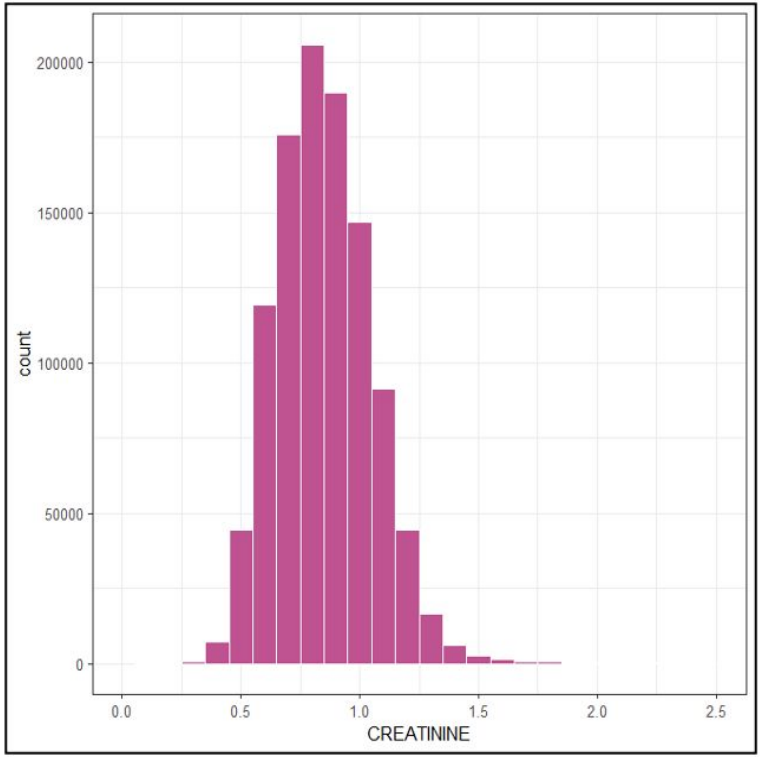


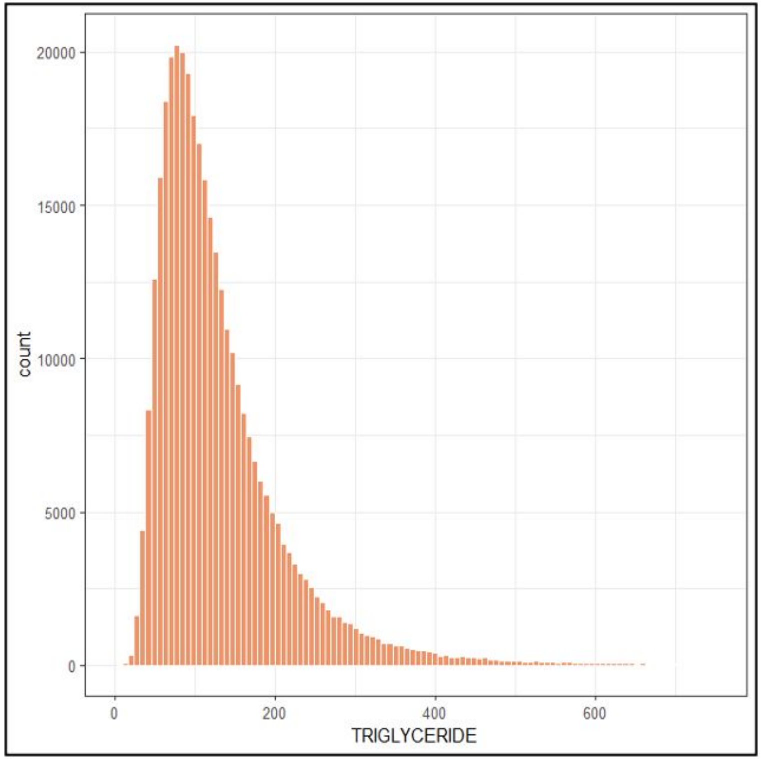






2019 Key variables





Methods and Data PreProcessing

- Utilized the following models from the SKLearn library
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier

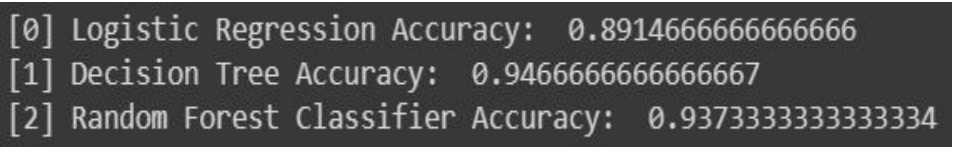
- Numeric variables scaled using StandardScaler

- Our Response Variable, SGOT_AST, was encoded from a numeric to yes/no depending on the value being above or below 40(0-40 being the normal level expected)

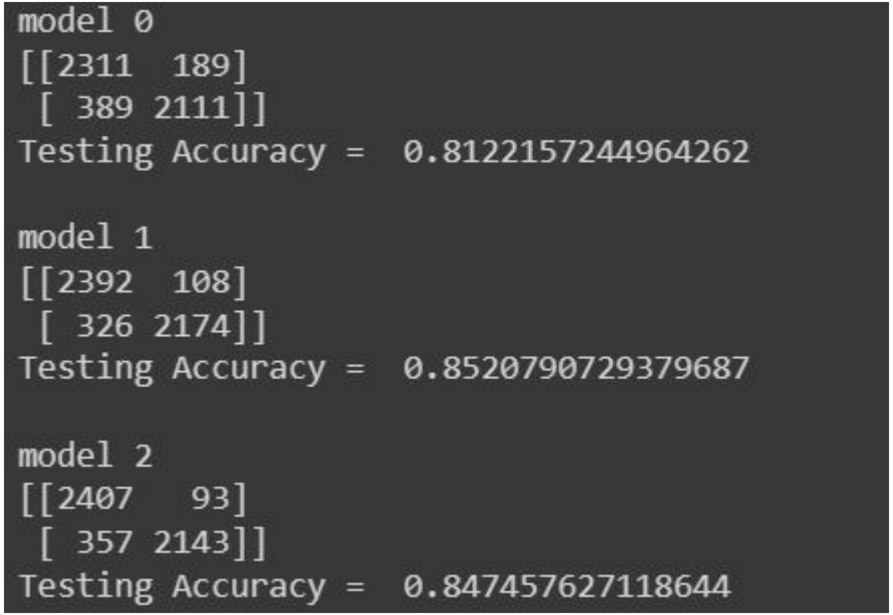
- Hyperparameter Tuning with GridSearchCV
  - Tol and C in Logistic Regression Model
  - Max_Depth and Criterion in Decision Tree Classifier Model
  - Max_Depth and N_estimators in Random Forest Classifier Model
- Training/Test Splits created with ratio of 75:25

Results

- 2009

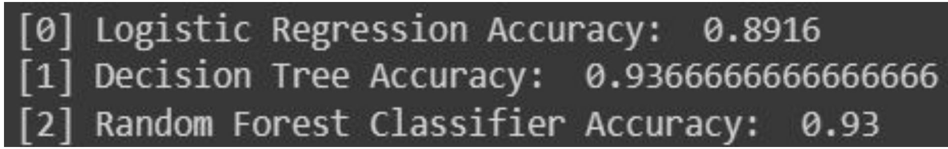


```
[0] Logistic Regression Accuracy:  0.8914666666666666
[1] Decision Tree Accuracy:  0.9466666666666667
[2] Random Forest Classifier Accuracy:  0.9373333333333334
```

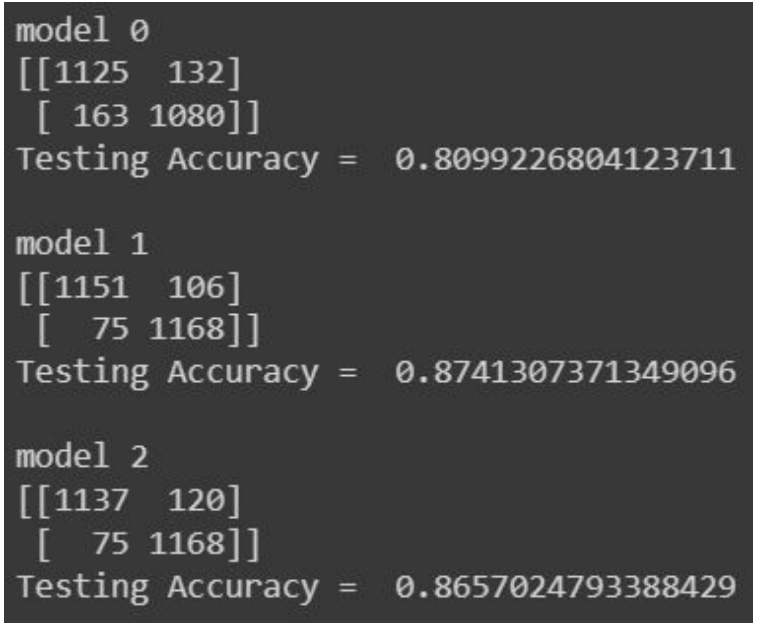


```
model 0
[[2311  189]
 [ 389 2111]]
Testing Accuracy =  0.8122157244964262

model 1
[[2392  108]
 [ 326 2174]]
Testing Accuracy =  0.8520790729379687

model 2
[[2407   93]
 [ 357 2143]]
Testing Accuracy =  0.847457627118644
```

2019



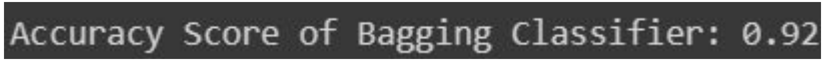
```
[0] Logistic Regression Accuracy:  0.8916
[1] Decision Tree Accuracy:  0.9366666666666666
[2] Random Forest Classifier Accuracy:  0.93
```



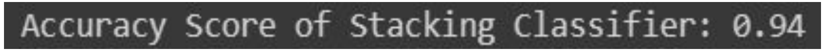
```
model 0
[[1125  132]
 [ 163 1080]]
Testing Accuracy =  0.809922680412371l

model 1
[[1151  106]
 [  75 1168]]
Testing Accuracy =  0.874130737134909б

model 2
[[1137  120]
 [  75 1168]]
Testing Accuracy =  0.8657024793388429
```

Ensemble Method Results

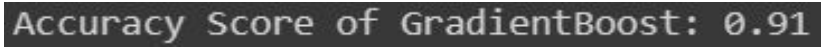
- 2009



```
Accuracy Score of Bagging Classifier: 0.92
```

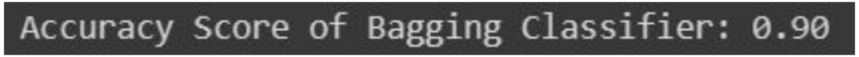


```
Accuracy Score of Stacking Classifier: 0.94
```

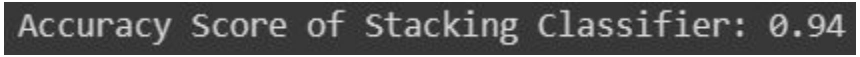


```
Accuracy Score of GradientBoost: 0.91
```

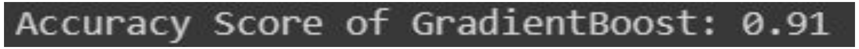
2019



```
Accuracy Score of Bagging Classifier: 0.90
```



```
Accuracy Score of Stacking Classifier: 0.94
```



```
Accuracy Score of GradientBoost: 0.91
```

Results

- Able to build several models that predicted the correct liver enzyme levels at an accuracy rate between 81% and 94%
  - We are able to accurately predict elevated SGOT_AST liver enzyme levels and thus possible liver problems at a high rate

- Utilized subsets of the data for each model to avoid long computation times with 1 million observations

Future Efforts

- Utilize more observations to create a more accurate predictive model

- Incorporate other indicators into the model

- Analyze liver enzyme levels for people of different countries

- Work to implement these models into a clinical setting to help notify doctors when a patient may be at risk of a liver problem due to heightened liver enzyme levels

THANK YOU FOR LISTENING
IF YOU HAVE ANY QUESTIONS I'M SURE GOOGLE WILL ANSWER THEM
makeameme.org

Thank you!