Natalie Bondurant, Dylan Deng, Ruiqi Wang, Huijie Lai
nbondurant@ucdavis.edu, dydeng@ucdavis.edu, rqqwang@ucdavis.edu, jaclai@ucdavis.edu
Dr. Horiguchi
STA 141A
12 December 2024

STA 141 Final Project

Executive Summary

This project's goal is to find the relationship between client-level variables and mental health outcomes using a dataset from Substance Abuse and Mental Health Services Administration(SAMHSA). The dataset includes detailed information of diagnoses for people who have used the mental health services in the year of 2022. There are three goals we are trying to achieve:

1. Predict the first mental health diagnosis using demographic variables such as age, gender, and marital status.
2. Using the trained prediction of these variables to predict the presence of specific mental health disorders, we are focused on trauma-related and depressive disorders.
3. Find gender differences in the prevalence of mental health disorders.

Since there were too many observations in the original dataset, we instead created a subset of 10,000 observations to finish our research. Across the research, we used logistic regression and cross-validation techniques to help us analyze the dataset.

Significant predictors for the first diagnosis include age, ethnicity, gender and marital status. Throughout our research, we found highly notable gender differences in mental health outcomes as we examine the third question.

These findings could be useful in the future to refine the prediction model. Furthermore, we can explore additional variables to improve the accuracy and put the prediction model in use in the study of mental health areas, which might improve the process of diagnosis for doctors.

Description of project, background, goals

In our project, we will be looking at a dataset that was originally prepared by the Substance Abuse and Mental Health Services Administration (SAMHSA). The main focus of this dataset is information regarding mental health diagnosis, the use of various mental health services, and client-level data, including demographics. The dataset was only collected from people who have used the mental health services, which includes demographics, national measures and substance use characteristics. There are a humongous amount of observations in the dataset, but with consideration of time consumption and accuracy, our group decided to use the first 10,000 patients' observations in our analysis.

Our project is guided by the following questions:

1. Can the first mental health diagnosis be predicted using client-level variables such as gender, employment status, and other demographic factors?
2. Is it possible to predict the presence or absence of a disorder based on client-level variables?
3. Are there significant gender differences in the prevalence or types of mental health disorders, as evidenced by bar charts and confidence intervals?

It has long been observed that certain mental health disorders are more prevalent among certain genders. For example, females are more frequently diagnosed with anxiety and depressive disorders while males are more frequently diagnosed with attention-deficit disorders and substance use disorders. With this project, we aim to help verify these claims as well as look more deeply into if and how other client-level factors might influence the prevalence of mental health disorders.

We are interested in these questions as they relate to the dataset and are concerned about modern people's mental health under certain conditions. We are dedicated to finding some relationship between mental health and other variables. Furthermore, we aim to use our findings to predict the outcome of patients' mental health. Therefore, our main goals are to see if we can predict mental health diagnoses using client-level variables using k-fold cross validation and also leave-one-out-cross-validation (LOOCV), then compare them to find out which method predicts the value better. Secondly, we are interested in predicting a specific kind of disorder based on client-level variables using logistic regression. Lastly, we want to discover any significant gender differences in the prevalence or types of disorder and visualize it with plots and graphs.

By fulfilling our goals and after calculation, we will be able to find relation between variables from the dataset and use the prediction model to predict any further data of patients to make the mental health treatment process easier and be more accurate. Moreover, a survey can be generated to ask patients to fill out the body conditions and uses our model to predict the outcome in order to alleviate the pressure of both patient and doctor. Knowing how various demographics may impact the risk of certain mental disorders could be useful for tailoring treatments and preventative measures towards certain groups of people. Obviously, there are infinitely many more factors that impact the risk of developing mental disorders, but if there are strong correlations between specific demographics and disorders, we may be able to initiate more specialized and customized programs to help people who struggle with mental health problems.

Description of dataset and data source:

We used a dataset collected in 2022 by the Substance Abuse and Mental Health Services Administration involving client-level data, use of mental health services, and mental health diagnoses. SAMHSA, part of the U.S. Department of Health and Human Services, is an organization aiming to educate people about mental health by making related data, services, and research more accessible. They combined two of the datasets collected by the states, the Mental Health Client-Level Data (MH-CLD) and the Mental Health Treatment Episode Data Set (MH-TEDS). As suggested by its name, MH-CLD focused heavily on participant-level variables such as gender, ethnicity, race, and location. The other dataset, MH-TEDS, emphasized mental-health related variables such as diagnoses, treatments, and types of state services used.

There were 6,957,919 of 40 variables, of which we used the first 10,000 rows to increase efficiency. Of these variables, one was used to identify participants which we omitted from our analysis, 38 were categorical variables, and one (AGE) was continuous. Variables were selected pertaining to our different research questions, and we tended to utilize the client-level data such as gender and race as well as the mental health diagnosis data. For example, to examine the relationship between client-level data and first diagnosis, we looked at the following client-level variables: age, ethnicity, race, gender, marital status, and employment status. We then used these variables to see if we could accurately predict what their first mental health diagnosis might be.

Our next question involved the variables age, ethnicity, race, gender, state, employment status, marital status, use of various inpatient services, and number of mental health diagnoses reported. From these patient-level variables, our goal was to see if we could predict the presence of specific types of disorders. For simplicity, we focused on trauma-related disorders and depressive disorders. Finally, our third question looked for significant gender differences in the prevalence or types of various disorders. As shown by Figure 1 below, there do visually appear to be some potentially significant differences in the proportion of males and females diagnosed with certain types of mental disorders.

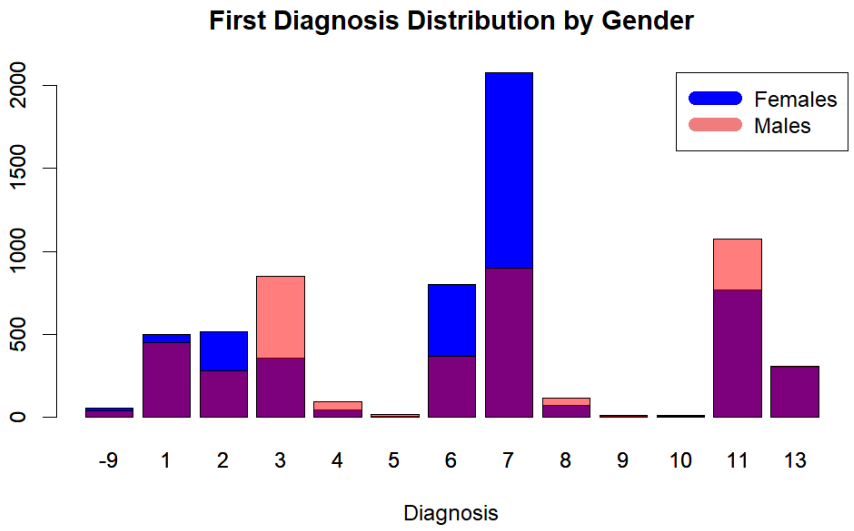


Figure 1: This bar chart of diagnosis plotted against frequency, with females' frequency displayed in blue and males' in pink, shows some gender differences in the frequency at which certain disorders are reported. -9 represents unknown diagnosis, 1 represents trauma-related disorders, 2 represents anxiety, 3 represents ADHD, 4 represents conduct disorders, 5 represents dementia disorders, 6 represents bipolar disorders, 7 represents depressive disorders, 8 represents oppositional defiant disorders, 9 represents pervasive developmental disorders, 10 represents personality disorders, 11 represents psychotic disorders, 12 represents substance use disorders, and 13 represents other disorders.

```
f <- dat[ which(dat$GENDER=="2"), ]
f
m <- dat[ which(dat$GENDER=="1"), ]
m
```

```
diagf <- table(f$MH1)
diagm <- table(m$MH1)

barplot(diagf,
        main = "First Diagnosis Distribution by Gender",
        xlab="Diagnosis",
        col="blue") #label x axis with diagnoses
barplot(diagm,
     col=rgb(1,0,0,0.5),
     add=TRUE)
legend("topright", legend=c("Females","Males"), col=c("blue","lightcoral"),lwd=10)
```

So, because our bar chart showed likely significant gender differences, we decided to test whether or not there are significant gender differences in the prevalence or types of disorders. For simplicity, in Figure 2 we report here our findings on depressive disorders, which displayed the largest gender differences.

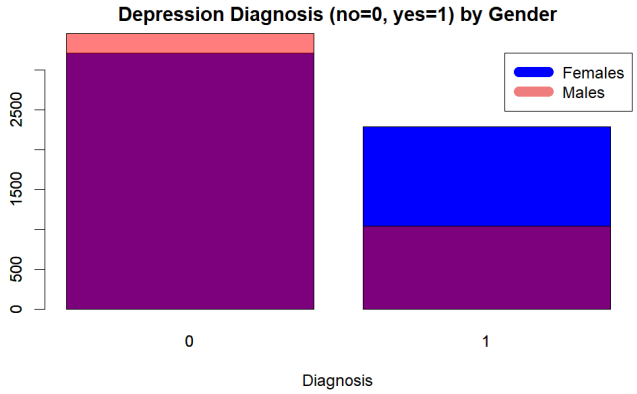


Figure 2: This bar chart shows the prevalence of depressive disorders in females (blue) versus males (pink). 0 represents the absence of a diagnosed depressive disorder, while 1 represents the presence of one. It shows that the proportion of people diagnosed with a depressive disorder is around two-thirds female and one-third male. Additionally, we see that just over 50% of people without a depressive disorder are male while less than 50% are female.

```
diagf <- table(f$DEPRESSFLG)
diagm <- table(m$DEPRESSFLG)
barplot(diagf,
        main = "Depression Diagnosis (no=0, yes=1) by Gender",
        xlab="Diagnosis",
        col="blue") #label x axis with diagnoses
barplot(diagm,
     col=rgb(1,0,0,0.5),
     add=TRUE)
legend("topright", legend=c("Females","Males"), col=c("blue","lightcoral"),lwd=10)
```

These bar charts showed us that it could be interesting to look more deeply into the gender gaps in the prevalence of specific types of mental disorders.

Limitations of the dataset are that each client was only able to report up to three diagnoses and that the population was limited to only people who had utilized mental health services in locations that report their findings to their state; the state mental health agencies, or SMHAs, are required to release patient information to the government within a 12-month period defined by date of admission.

Below is a chart listing the variables in our dataset that we focused on during our analysis and their corresponding labels.

Variable	Label	Variable	Label
EDUC	Education	AGE	Age at time of data collection (in 2022)
ETHNIC	Hispanic or Latino origin (ethnicity)	RACE	Race
EMPLOY	Competitive employment status (aged 16 years and older) at discharge or end of the reporting period	LIVARAG	Residential status — at discharge or end of reporting period
GENDER	Sex (Male, Female, or Unknown)	MH1	First mental health diagnosis

MARSTAT	Marital status	DEPRESSFLG	Depressive disorder reported
TRAUSTREFLG	Trauma- and stressor-related disorder reported	NUMMHS	Number of mental health diagnoses reported

Methodology

Question 1

We used 10-fold cross-validation (10-fold CV) with GLM to assess the accuracy of predicting MH1 based on variables such as AGE, ETHNIC, EDUC, and others. In this method, the dataset was randomly divided into 10 equally sized subsets. For each fold, one subset was used as the test set, and the remaining nine were used as the training set. This process was repeated 10 times, ensuring that every observation was used for both training and testing exactly once. Because we were interested in randomly splitting our dataset into a training set and a validation set in order to assess the performance of the fit of the model, cross validation was an appropriate method. We used k-fold cross validation because it was less computationally intensive and because it has a smaller variance than LOOCV because LOOCV has smaller bias.

To predict the value of MH1, we used linear regression. Since MH1 is a continuous response variable, we specified family = gaussian in the glm() function to model the relationship between the predictors and the response. The fitted model was then used to generate predictions for the test set in each fold. The predictions across all folds were aggregated and compared to the actual values of MH1 to evaluate the model's performance.

To visualize and assess the prediction errors, we plotted a boxplot and a histogram of the differences between the actual and predicted values of MH1. Additionally, we used the summary() function to obtain a numerical summary of the prediction errors, which helped us better understand the distribution and accuracy of the predictions.

```
mydata <- read.csv("mhcld_puf_2022.csv",nrows=10000)
```

```
mydata.q1 <- mydata1[,c("AGE","ETHNIC","EDUC","RACE","GENDER",'MARSTAT','EMPLOY','MH1')]
View(mydata.q1)
fit.q1 <- glm(MH1 ~ ., family = gaussian, data = mydata.q1)
summary(fit.q1)
```

```
predict(fit.q1, type = "link")
```

```
set.seed(123)
fold <- sample(1:1000, replace = F)

mydata.q1 <- mydata.q1[fold,]
mydata.q1$fold.index <- rep(1:10, each = 100)
mydata.q1$tenfoldcv <- rep(NA, 1000)

for (i in 1:10) {

  data.train <- mydata.q1[-which(mydata.q1$fold.index == i),]
  data.test <- mydata.q1[which(mydata.q1$fold.index == i),]

  fit.temp <- glm(MH1 ~ AGE + ETHNIC + EDUC + RACE + GENDER + MARSTAT + EMPLOY, family = gaussian, data = data.tra
in)
  mydata.q1$tenfoldcv[which(mydata.q1$fold.index == i)] <- predict(fit.temp, newdata = data.test)

}

boxplot(mydata.q1$MH1 - mydata.q1$tenfoldcv)
```

```
hist(mydata.q1$MH1 - mydata.q1$tenfoldcv)
```

```
summary(mydata.q1$MH1 - mydata.q1$tenfoldcv)
```

Question 2

To see if we could accurately predict the presence or absence of a specific diagnosis based on other variables in the dataset, we used logistic regression. We are using variables Age, Ethnic, Education, Race, Gender, Marstat and Employment to predict our dependent variable DEPRESSFLG, which represents whether or not a depressive disorder was reported in each client. Those independence variables are selected to investigate their fluence on the binary outcome of DEPRESSFLG.

Since our response variable was binary (with 0 representing not having the disorder and 1 representing having the disorder), we used logistic regression to model the probabilities of binary responses between 0 and 1, where values less than 0.5 were coded as 0 and values greater than or equal to 0.5 were coded as 1. For simplicity, we used DEPRESSFLG as our binary response variable, so 0 represented the absence of a depressive disorder diagnosis and 1 represented the presence of a depressive disorder diagnosis. Using the glm() function for logistic regression, we specified that DEPRESSFLG was a binary response by including the argument "family=binomial."

In the code, we are using a logistic regression model to predict the binary outcome by using the independence variables. The regression was performed using the function 'glm', which stands for Generalized Linear Model, and binomial family to account for the binary outcome.

We started off our code with model training without any validating. Prediction for training data was generated with a probability threshold of 0.5 that is used to classify the outcomes. This training model was used to compare with later predicted outcomes with true values.

Prediction by using logistic regression:

```
mydata.q2 <- mydata1[,c("AGE","ETHNIC","EDUC","RACE","GENDER",'MARSTAT','EMPLOY','DEPRESSFLG')]
View(mydata.q2)
fit.q2 <- glm(DEPRESSFLG ~ ., family = binomial, data = mydata.q2)
summary(fit.q2)
```

```
predicted.q2 <- predict(fit.q2, type = "response")
predicted.q2 <- ifelse(predicted.q2 > 0.5, 1, 0)

table(mydata.q2$DEPRESSFLG, predicted.q2)
```

Lastly, in order to obtain a reliable prediction of our model, the Leave-one-out cross-validation (LOOCV) method was conducted. For each run of LOOCV, one observation was held out as the test set and others as a training set. Then the logistic regression model was fitted to the training data and predicted the outcome, this process is repeated for all our selected 1,000 patients. And the outcome will be converted to binary using the cutoff of 0.5 then comparing it with true values. The comparison is performed by confusion matrices in order to get the accuracy of prediction.

Prediction by using LOOCV

```
response.q2 <- rep(NA, nrow(mydata.q2))

for (i in 1:nrow(mydata.q2)) {

  data.train <- mydata.q2[-i,]
  data.test <- mydata.q2[i,]

  fit.temp <- glm(DEPRESSFLG ~ ., family = binomial, data = data.train)
  response.q2[i] <- predict(fit.temp, type = "response", newdata = data.test)
}

mydata.q2$loocv <- response.q2
View(mydata.q2)

mydata.q2$loocv <- ifelse(mydata.q2$loocv > 0.5, 1, 0)
table(mydata.q2$DEPRESSFLG, mydata.q2$loocv)
```

Question 3

Prior to conducting the chi-square test for gender differences in the prevalence of disorders (DEPRESSFLG), we began with a basic exploratory data analysis. We started by selecting the relevant variables (GENDER and DEPRESSFLG) and reviewing them through the function view() and table(). These gave us the data distribution and allowed us to confirm the variables.

For the analysis itself, we relied on a traditional statistical inference method, chi-square test((chisq.test(table())), to evaluate whether there was a statistically significant association between gender and the distribution of the disorder variable. It provided a straightforward and interpretable measure of the relationship between these categorical variables.

```
mydata.q3 <- mydata1[,c('GENDER', 'DEPRESSFLG')]
View(mydata.q3)

table(mydata.q3)
```

```
chisq.test(table(mydata.q3))
```

```
result.q3 <- as.data.frame(table(mydata.q3))
result.q3 <- result.q3[c(1,3,2,4),]
result.q3$percentage <- rep(NA, 4)
result.q3$percentage[c(1,2)] <- paste0(round(result.q3$Freq[c(1,2)]/sum(result.q3$Freq[c(1,2)])*100), "%")
result.q3$percentage[c(3,4)] <- paste0(round(result.q3$Freq[c(3,4)]/sum(result.q3$Freq[c(3,4)])*100), "%")
View(result.q3)
```

In contrast, for comparison purposes, we also conducted the same analysis using the variable TRAUSTREFLG instead of DEPRESSFLG. This allowed us to investigate whether gender differences varied between trauma and stressor related disorders and depressive disorders.

```
mydata.q3 <- mydata1[,c('GENDER', 'TRAUSTREFLG')]
View(mydata.q3)

table(mydata.q3)
```

```
chisq.test(table(mydata.q3))
```

```
result.q3 <- as.data.frame(table(mydata.q3))
result.q3 <- result.q3[c(1,3,2,4),]
result.q3$percentage <- rep(NA, 4)
result.q3$percentage[c(1,2)] <- paste0(round(result.q3$Freq[c(1,2)]/sum(result.q3$Freq[c(1,2)])*100), "%")
result.q3$percentage[c(3,4)] <- paste0(round(result.q3$Freq[c(3,4)]/sum(result.q3$Freq[c(3,4)])*100), "%")
View(result.q3)
```

Analysis and Findings

Question 1

From the result calculated by R, we can see that predictors Age, Ethnic, Gender and MARSTAT are significant since their p-value is smaller than 0.05, which is the significance level. Where other variables such as Educ, Race and Employ would have less or almost no impact on Mental Health 1.

The box plot clearly shows the distribution of difference between the actual value and predicted values using the 10-fold cross validation. From the plot, we can find that the median of the errors is close to 0, which indicates the prediction of the 10-fold method is unbiased. And most of the prediction errors are small and numbers are around 0, suggesting that the predicting model did well on average. And clearly, there are some outliers that indicate the model cannot perfectly predict our value.

The histogram is illustrating the frequency of prediction errors. Where it is showing a shape close to symmetric, a little bit skewed left. From this shape, we can say the errors of prediction do not systematically overpredict or underpredict.

Above all, our predictions are close enough to our actual values that we can possibly use our model to predict MH1's outcome by using other variables. However, there are still some outliers shown on boxplot, some improvement can be made to adjust those outliers, for example, get rid of those variables that are not significant enough to affect our prediction, such as Educ, Race and Employment.

```
glm(formula = MH1 ~ ., family = gaussian, data = mydata.q1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17767    1.28351   2.476  0.01346 *
AGE          0.37040    0.03191  11.607  < 2e-16 ***
ETHNIC       0.57866    0.26519   2.182  0.02934 *
EDUC         0.08435    0.05500   1.534  0.12545
RACE        -0.16464    0.10439  -1.577  0.11508
GENDER      -0.53941    0.24125  -2.236  0.02558 *
MARSTAT     -0.22772    0.08271  -2.753  0.00601 **
EMPLOY      -0.02176    0.01963  -1.109  0.26789
```

Figure 1. Summary of the Gaussian Generalized Linear Model, showing the relationship between variables and dependent.

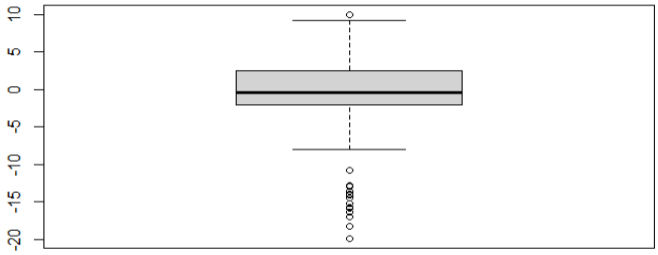
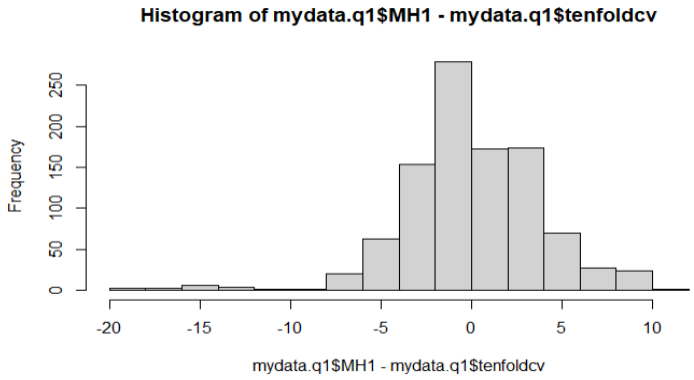


Figure 2. Box plot generated by prediction outcomes and it is showing the error of predicting value and real value.



Question 2

Using the summary function on our logistic regression, we found that the following client-level variables contributed significantly to the presence or absence of a depressive disorder diagnosis: age, education, gender, employment status, use of state psychiatric hospital services, use of other inpatient psychiatric services, residential status, and the number of mental health diagnoses.

We then created an accompanying confusion matrix to determine the accuracy of our logistic regression. Type I error was defined as having the model predict the presence of a disorder when the disorder was absent, Type II error was defined as having the model predict the absence of a disorder when the disorder was present, true positive was defined as the model correctly predicting the presence of a disorder, and true negative was defined as the model correctly predicting the absence of a disorder. Figure 3 below displays a table of these values.

		True Class	
		- or Null	+ or Non-null
Predicted	- or Null	True Negative: 5,937	False Negative: 740
Class	+ or Non-null	False Positive: 2,280	True Positive: 1,043

Figure 3: A confusion matrix of the logistic regression shows the number of values for true negative, false negative, false positive, and true positive.

From Figure 3, we calculated the accuracy of our model to be 69.8%, meaning the logistic regression model correctly predicted the presence or absence of a disorder about 70% of the time using the specified client-level variables. As expected, the number of false positives was fairly high; the prevalence of depressive disorders is considerably lower than 50%, meaning our model would likely predict more positives than true number of incidences. We also know from our model used in Question 1 that certain variables are more strongly correlated with the presence of some disorders such as gender and age.

Question 3

We conducted a chi-square test to examine whether gender is associated with Trauma- and stressor-related disorder (DEPRESSFLG). The resulting p-value was approximately 1.363e-09. This extremely low p-value strongly suggests that gender is significantly associated with the prevalence of depressive disorders. The stark difference in significance level between the two variables that we tested highlights the importance of examining the multiple mental health outcomes, as gender-related differences may be more pronounced in certain disorders than others.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(mydata.q3)
## X-squared = 36.721, df = 1, p-value = 1.363e-09
```

In contrast, we tried this method on a different variable,TRAUSTREFLG, and found that the resulting p-value was approximately 0.05351, which is slightly above the conventional threshold of 0.05. This indicates that we did not find a statistically significant difference in the prevalence of this disorder between genders based on our sample.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(mydata.q3)
## X-squared = 3.728, df = 1, p-value = 0.05351
```

Conclusions

Using the information from the data Substance Abuse and Mental Health Services Administration, the project explores the relationship between factors, such as age, gender, material status, mental health, and more. This study focused on following points:

1. Predicting clients' most probable first mental health diagnoses using client-level variables such as gender and employment, then noting which of these variables contributed more significantly to the diagnosis
2. Predicting the presence or absence of a disorder using client-level data
3. Examining gender differences in disorders

With logistic regression and cross-validation methods, like k-fold and leave-one-out cross-validation (LOOCV), we found that age, ethnicity, gender, and marital status were significant predictors for the first mental health diagnosis (MH1). These factors have p-values smaller than the significance level of 0.05. Other variables, such as education, race, and employment, had less impact on predictions. The model, predicted with these factors, achieved an accuracy of 69.8% when predicting depressive disorders but displayed a high false-positive rate, indicating it often predicted a disorder when there wasn't one.

When examining gender differences in causing depressive disorders, the chi-square test is used. The result shows that depressive disorders are significantly more common in women than men. In comparison, there are no significant differences found for trauma- related disorders between the genders. This highlights the importance of exploring each condition separately. The result also emphasized the role of demographic factors in understanding the prevalence and types of mental health disorders.

Through our analysis, we achieved significant progress toward these goals. Our use of LOOCV to predict clients' first mental health diagnoses was shown to be moderately accurate, which could be progress towards making a more comprehensive model that incorporates both more and different client-level variables in order to better predict their first diagnoses. This could serve as a method of screening clients to look for risk factors for different disorders, and this information could be used to educate people on preventative measures, symptoms of the disorder, and treatment options and how to access them.

Looking more closely at specific disorders, our logistic regression model also gave fairly accurate predictions of whether or not a disorder would be present given a client's patient-level data. So, similar to the LOOCV model, the logistic regression model could be a useful tool for predicting the presence or absence of specific disorders. Other variables that could be useful in predicting this could include looking in clients' family histories for the presence of the disorder since many mental disorders have at least some genetic component. The dataset we used in this analysis did not include such a variable, but future research in a similar area may be an interesting way to better understand and predict when certain mental disorders might manifest, which could be used to help provide at-risk individuals with the proper education and resources to reduce their chances of developing the disorder or minimize any symptoms of the disorder.

Because the current understanding of mental disorders and our initial analysis show that gender can play a large role in impacting the types of mental disorders that are diagnosed in clients. For example, we found that depression is much more commonly diagnosed in women while trauma disorders are not diagnosed at significantly different rates in women versus men. Additionally, our analysis cannot tell us if these gender differences are due to different rates of prevalence of these disorders or different rates of diagnosis. So, further analysis using other datasets may be helpful in determining the causes for this gender gap.

With this knowledge and more, we can aim to improve the mental health treatment process and preventative measures by tailoring mental health services and education towards

people who may be at a higher risk of having certain mental disorders. As a result, hopefully more people will be able to better understand their mental health, how to take care of themselves, and how to receive any help they may need.