**Mental Health Patterns and Living Situations in New York State**

|  |  |
| --- | --- |
| Prepared for: | Professor Ning Rui |
| Prepared by: | Dania Salman  Emma Francis  Heng-Jui Chu |

December 11, 2025

**Table of Contents**

[**Introduction** 3](#_Toc216356045)

[**SMART Questions** 5](#_Toc216356046)

[**Data Cleaning** 6](#_Toc216356047)

[**SMART Question 1** 6](#_Toc216356048)

[**Filtering variables** 6](#_Toc216356049)

[**Correlation Heatmap** 8](#_Toc216356050)

[**Logistic Regression** 8](#_Toc216356051)

[**Random forest model** 10](#_Toc216356052)

[**Feature importance model** 10](#_Toc216356053)

[**Logistic regression (Balanced the class weight)** 11](#_Toc216356054)

[**Random Forest (Balanced the class weight)** 12](#_Toc216356055)

[**Smart question 1 Conclusion** 13](#_Toc216356056)

[**SMART Question 2** 13](#_Toc216356057)

[**Correlation Heatmap** 15](#_Toc216356058)

[**Logistic Regression Model (Balanced class weights)** 16](#_Toc216356059)

[**Random Forest Model (Balanced class weights)** 17](#_Toc216356060)

[**Feature importances for Random Forest Model (Balanced class weights)** 18](#_Toc216356061)

[**Smart question 2 conclusion** 19](#_Toc216356062)

[**SMART Question 3** 19](#_Toc216356063)

[**Data Description** 20](#_Toc216356064)

[**Correlation test: Cramér’s V matrix** 20](#_Toc216356065)

[**Multinomial Logistic Regression** 21](#_Toc216356066)

[**Random Forest** 22](#_Toc216356067)

[**Random Forest & SMOTE** 22](#_Toc216356068)

[**Random Forest & SMOTE (Without minority class)** 23](#_Toc216356069)

[**Feature Importance for Random Forest & SMOTE (Without minority class)** 24](#_Toc216356070)

[**Smart question 3 conclusion** 25](#_Toc216356071)

[**Conclusion** 25](#_Toc216356072)

[**Recommendations** 26](#_Toc216356073)

[**References** 26](#_Toc216356074)

# **Introduction**

Mental health and housing stability are ongoing crises in New York state. Many New York residents have a mental illness which can be triggered by financial stress, limited health insurance coverage, and overall less support. The Substance Abuse and Mental Health Services Administration reported that roughly 19.5 percent of adults who live in New York had a mental illness, with about 3.7 percent having a serious mental illness in 2019 (SAMHSA, 2019). Ettman et al. (2013) found in their systematic review that financial instability is a consistent and prevalent predictor of depression across manifold populations. Since New York has one of the highest costs of living, it is important to examine mental health patterns and living situations in the state.

New York, like many states in the nation, is facing a housing crisis. With high housing costs and a limited supply, it is growing increasingly difficult for New York State residents to have housing stability. Urban areas, New York City especially, have some of the highest housing costs in the country, while New York’s more rural and suburban areas face a lack of housing supply (Office of the State of New York Comptroller, 2019). A struggling housing market can catalyze/worsen stress, financial strain, making individuals more vulnerable to mental illness. Economic insecurity and stress are linked to higher rates of depression among other mental illnesses (Tanarsuwongkul et al., 2025). In fact, Mason et al. (2013) proves that increasing financial struggles with housing affordability damages individuals’ mental health. Mason et al. (2013) employed longitudinal evidence and accounted for demographic and socioeconomic variables when proving this hypothesis.

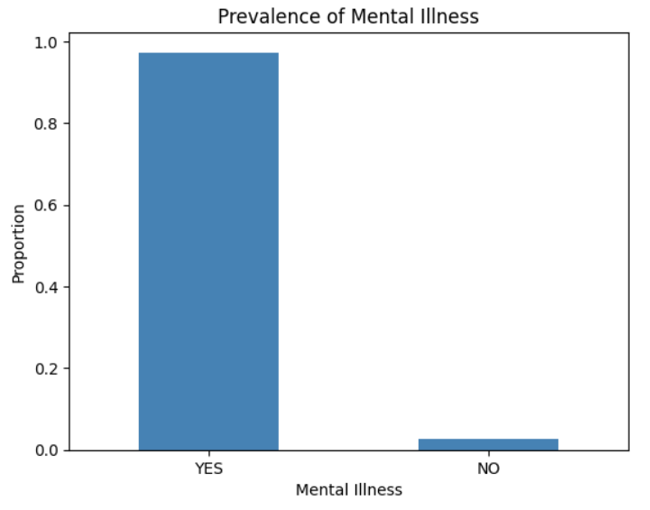
We chose to examine mental health patterns and living situations in New York state because New York is facing some of the highest costs of living in the nation and a struggling housing market. We are curious to see how it affects New York residents’ livelihoods. The existing literature examines how mental health patterns and living situations affect people in general, but no one examines this on a state-level. So, we aim to examine these variables in New York to help be able to inform policy.

**Data Description**

In this project, we utilize the 2019 Patient Characteristic Survey (PCS) for New York state to analyze relationships between demographic, socioeconomic, housing, and mental health variables. The PCS dataset provides comprehensive information on these issues since it includes information on a variety of mental health, housing stability, demographic, and socioeconomic variables. It is important to note that the PCS reports that over 90 percent of patients claimed to have some kind of a mental illness. The high prevalence of those who have a mental health illness reflects that the sampled group is part of a clinical population that is already receiving healthcare services.

Although the PCS consists of all types of patients, some survey participants are already receiving mental health treatment. Furthermore, mental illness commonly results with physical health conditions such as diabetes, obesity, chronic pain, among others (Moussavi et al., 2007; Katon, 2011). When patients seek physical health treatments, mental illnesses are sometimes also tested for with screenings. Focusing on a clinical population offers invaluable insights into how demographic, socioeconomic, mental health, and housing conditions interact, informing policies to improve mental health outcomes for patients in New York.

Because over 90 percent of individuals included in the PCS reported having a mental health diagnosis, our dataset has a class imbalance, which we will address in our statistical modeling.



# **SMART Questions**

We focused on the relationships between the prevalence of mental illness and demographic, socioeconomic, and housing stability variables. We picked these questions to gain further insight into mental health and housing situations in New York. The questions we examined are as follows:

1. How well can demographic characteristics predict whether an adult has a diagnosed mental illness in New York State using PCS 2019 data?
2. To what extent can socio-economic factors, particularly educational attainment, predict mental health diagnosis among adults in New York State using PCS 2019, and which variables are the strongest predictors?
3. How effectively can employment and related socioeconomic variables predict an individual’s living situation category in New York State using PCS 2019?

# **Data Cleaning**

We first subsetted and filtered our data using column names based on each SMART question. For example, the first question was filtered to only use the necessary columns that included mental health illness and demographic variables. Since PCS data is “YES/NO/UNKNOWN” data, we converted “UNKNOWN” values into missing values and then removed the missing values. We then encoded the categorical variables to be able to run statistical models. After data cleaning, we had 146,737 observations.

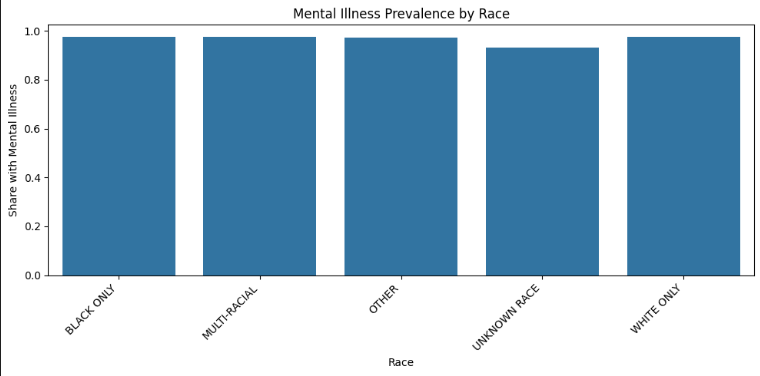
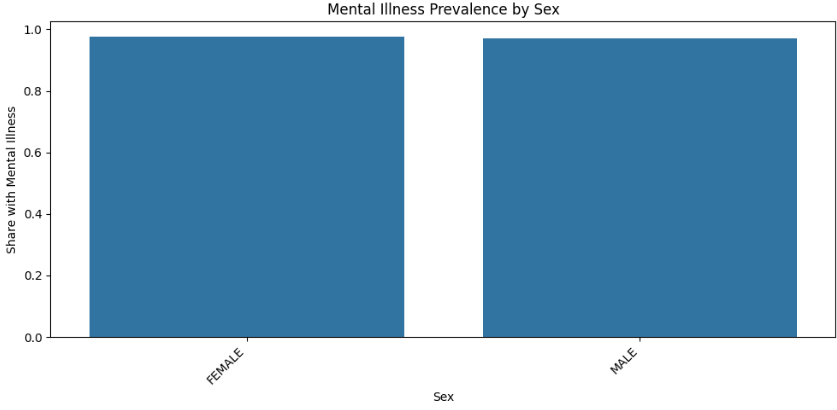
# **SMART Question 1**

How well can demographic characteristics predict whether an adult has a diagnosed mental illness in New York State using PCS 2019 data?

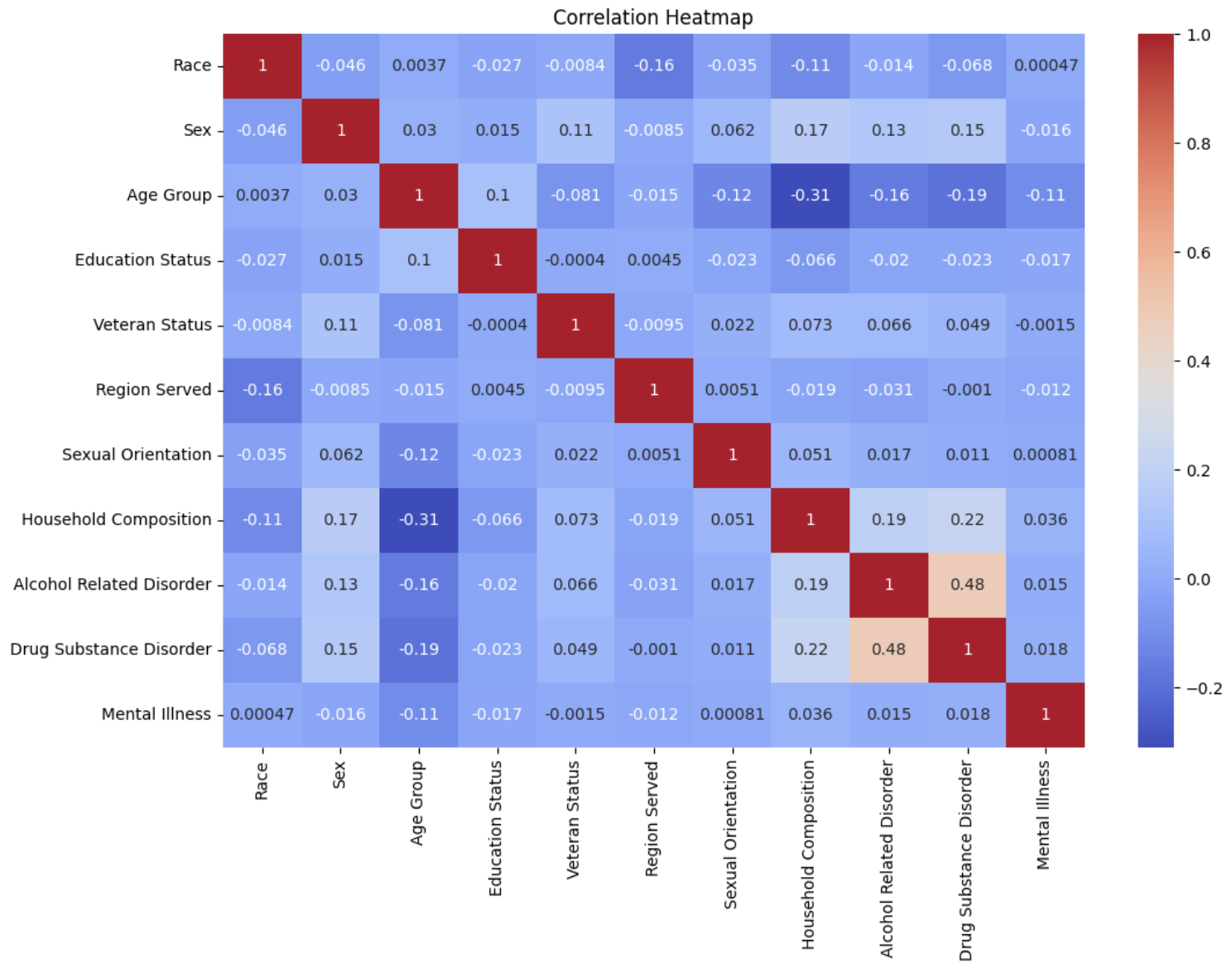
## **Filtering variables**

For SMART question 1, we filter the dataset out for relevant variables: race, sex, age group, education status, veteran status, region served, sexual orientation, household consumption, alcohol related disorder, and drug substance disorder. Since we have a class imbalance in the prevalence of mental illness, we continued a data exploration to see if the prevalence of mental health differed among the demographic variables. However, the prevalence of mental health is nearly equal among all groups in all demographic variables. Examples of mental illness prevalence among sex and race are shown below.

Although PCS has a variety of variables, it does not contain specific mental illness variables other than whether the respondent has a mental illness or a severe mental illness. So, we cannot test for a specific mental illness like depression. There is still an extreme class imbalance for severe mental illness, so we will use mental illness as our dependent variable.



## **Correlation Heatmap**

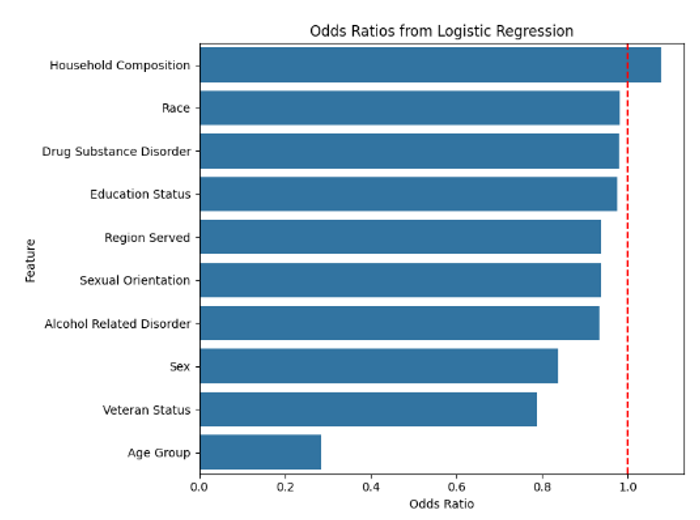
From these graphs, it is evident that mental illness does not differ significantly between sex and gender. We then examine the correlations between the variables. 

As seen from the correlation plot, the highest correlation among the variables is 0.48 which is considered low. So, no variables need to be dropped since multicollinearity does not need to be addressed. We then run a logistic regression.

## **Logistic Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **No Mental Illness** | 0.00 | 0.00 | 0.00 | 1034 |
| **Mental Illness** | 0.97 | 1.00 | 0.99 | 38670 |
| **Accuracy** |  |  | 0.97 | 39704 |
| **Macro Average** | 0.49 | 0.50 | 0.49 | 39704 |
| **Weighted Average** | 0.95 | 0.97 | 0.96 | 39704 |

The logistic regression highlights the class imbalance of the data. The accuracy of the model appears to be 97%. However, the model does not identify those who do not have a mental illness with both the precision and recall being 0 for “No Mental Illness”. The model has an AUC of 0.67, demonstrating that the logistic regression had modest discriminative ability. It is closer to a random fit since its AUC is only 0.17 away from 0.5. So, the current model is not suitable for the data due to the class imbalance. The odds ratios were then calculated and are displayed below:



From the results of the odds ratio, it is clear that cohabitating with others minimally increases the odds of having a mental illness. Also, being an adult decreases the odds of having a mental illness. This is a particularly interesting finding since mental illness tends to develop as one gets older, especially with financial stress. It is interesting that the other variables have small impacts on the odds of having a mental illness. From this model, it is clear that the independent variables do not explain sufficient variation in mental illness prevalence, likely due to the class imbalance. We then ran a random forest model on the data:

## **Random forest model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **No Mental Illness** | 0.64 | 0.05 | 0.09 | 1034 |
| **Mental Illness** | 0.98 | 1.00 | 0.99 | 38670 |
| **Accuracy** |  |  | 0.97 | 39704 |
| **Macro Average** | 0.81 | 0.52 | 0.54 | 39704 |
| **Weighted Average** | 0.97 | 0.97 | 0.96 | 39704 |

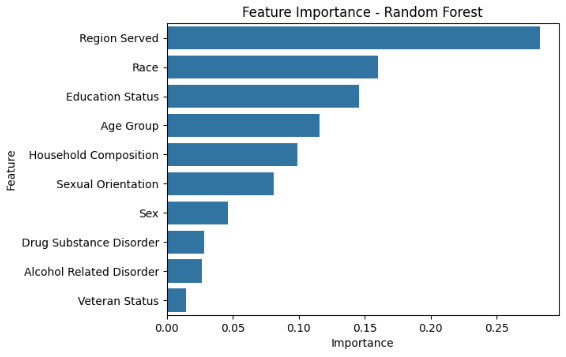
The Random Forest model also had an accuracy of 97% and an AUC of 0.69, again having some

discriminative ability. This model also suffers from the class imbalance, having a recall of 0.05

for individuals without a mental illness. This model is extremely biased toward those who do have a mental illness.

## **Feature importance model**

We then ran a feature importance model, finding that region served, education status, and race are the most significant predictors of mental illness; sexual orientation, household composition, and age group have some importance on the model; and sex, alcohol related disorder, drug substance disorder, and veteran status do not contribute much to the model's predictions. The results are displayed below.

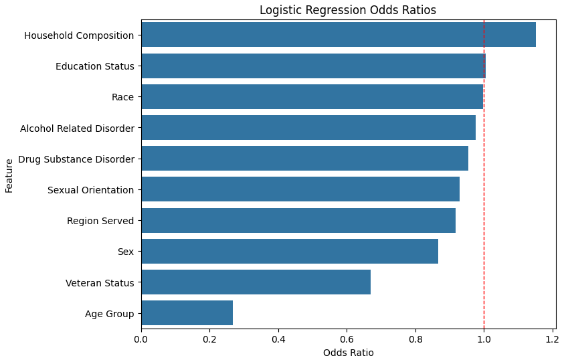


## **Logistic regression (Balanced the class weight)**

To account for the class imbalance, we balanced the class weight in another specification of the logistic regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **No Mental Illness** | 0.06 | 0.46 | 0.10 | 1034 |
| **Mental Illness** | 0.98 | 0.81 | 0.89 | 38670 |
| **Accuracy** |  |  | 0.80 | 39704 |
| **Macro Average** | 0.52 | 0.63 | 0.49 | 39704 |
| **Weighted Average** | 0.96 | 0.80 | 0.86 | 39704 |

Now that the class imbalance is accounted for in the model, a decent number of individuals who do not have a mental illness are identified. The model accuracy is now 76%, but it identifies those without a mental illness. The demographic/independent variables have modest predictive power now. We then calculated the odds ratios. The results are displayed below:



We got similar results before we balanced the class in the model. Cohabitating with others is associated with an increased odds of having a mental illness, and being an adult is associated with a lower chance of having a mental illness. Now, not being a veteran is associated with a lower chance of having a mental illness, which makes sense because non-veterans are less susceptible to developing mental health illnesses like post-traumatic stress disorder. However, most variables still have small impacts on mental illnesses.

## **Random Forest (Balanced the class weight)**

We then run a random forest model again, with the balanced class weight in the model. The random forest model now has a 75 percent accuracy and an AUC of 0.68. Although the accuracy did decrease, more individuals who do not have a mental illness are identified. The recall for those without a mental illness is now 0.54 which substantially improved from 0.05. The results are shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **No Mental Illness** | 0.05 | 0.54 | 0.10 | 1034 |
| **Mental Illness** | 0.98 | 0.75 | 0.85 | 38670 |
| **Accuracy** |  |  | 0.74 | 39704 |
| **Macro Average** | 0.52 | 0.64 | 0.47 | 39704 |
| **Weighted Average** | 0.96 | 0.74 | 0.83 | 39704 |

The feature importances were re-run with the balanced class weight. Each variable’s importance is roughly the same as before. Region served, education status, and age group are the most influential predictors of mental illness. And, same as before, drug substance disorder, alcohol related disorder, and veteran status contribute the least to the model.

## **Smart question 1 Conclusion**

Overall, when class imbalance is accounted for, a decent amount of those who do not have a mental illness are identified. Across both models, cohabitating with others increases the odds of having a mental illness, and being older reduces the odds of having a mental illness. Furthermore, region served, education status, age group, and race contribute the most to the models, and drug substance disorder, alcohol related disorder, and veteran status contribute least to the models. This suggests that age, can provide some insight into mental health status among adults in New York, but the class imbalance must be considered.

# **SMART Question 2**

For SMART question 2 we began by filtering the dataset to include the following socioeconomic variables

*Predicting variables:*

Living situation, Household Composition, Employment Status, Number of hours worked each week, Education status, Special Education services, Mental Illness, SSI Cash Assistance, SSDI Cash Assistance, Veterans Disability Benefits, Veterans Cash Assistance, Public Assistance Cash Program, Other Cash Benefits, Medicaid and Medicare Insurance, No Insurance, Unknown Insurance Coverage, Medicaid Insurance, Medicaid Managed Insurance, Private Insurance, Child Health Plus Insurance, Other Insurance, Criminal Justice Status,

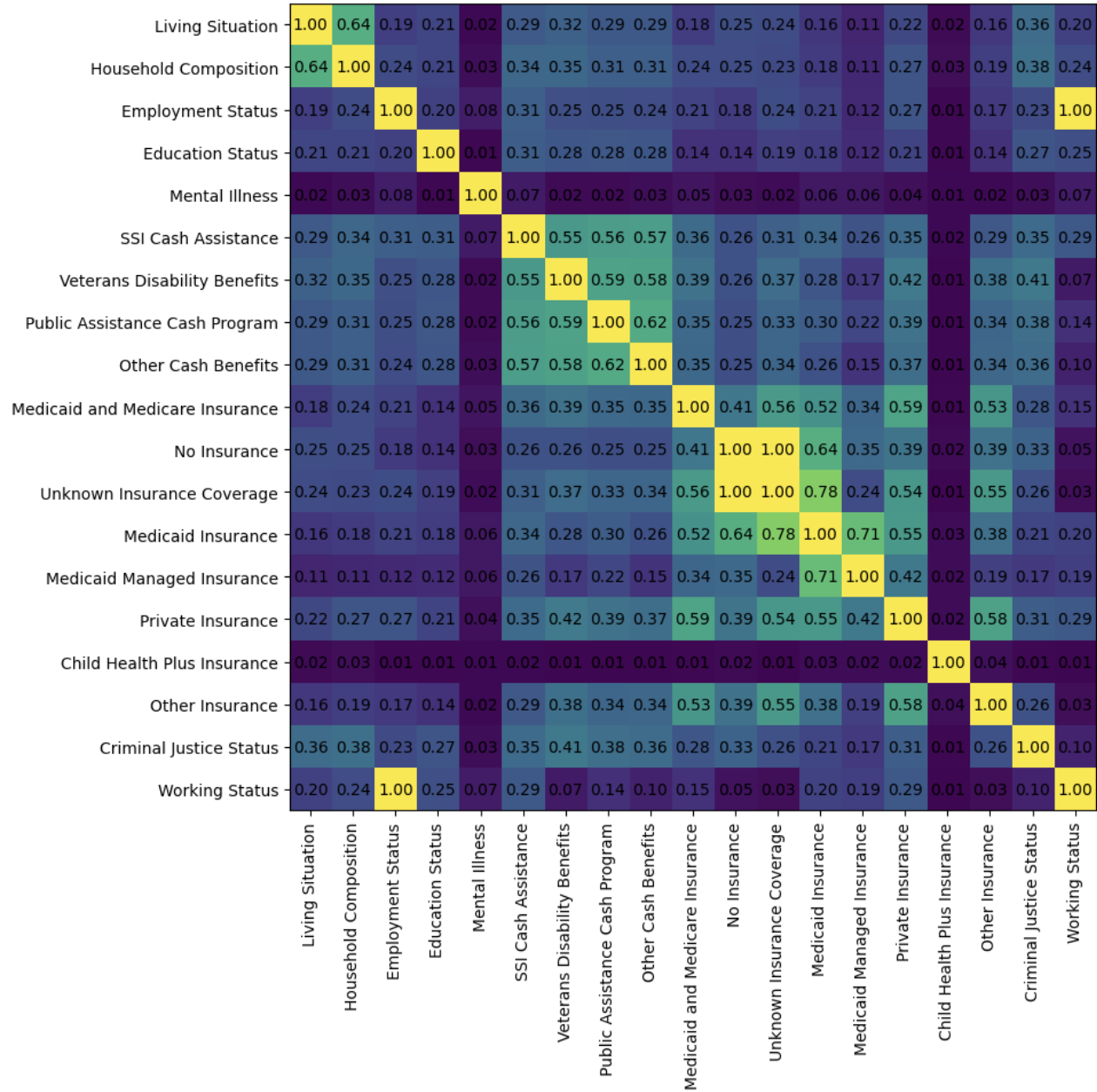
*Target Variable:* Mental Illness

After selecting these variables, we further filtered the data to include only entries where the Age Group was “ADULT”.

The following steps were then taken to reduce the number of predicting variables and reformat them to increase predictive power:

1. Check for the distribution of variables:
   1. Drop Special Education Services since it has no predictive power, and 99% of the responses are “NOT APPLICABLE”.
   2. Merge rare categories in Education Status, Number of Hours Worked Each Week, Veterans Cash Assistance, and Child Health Plus Insurance.
   3. Rename Number Of Hours Worked Each Week to Working Status.
2. Run Chi-squared tests between the remaining predicting variables and the target variable:
   1. Drop Veterans Cash Assistance since this was the only variable where p-value was not statistically significant.
3. Check Correlation between predicting variables to avoid multicollinearity when building models:
   1. Drop Household Composition, Employment status, Other Insurance, Private Insurance, Medicaid Insurance, Unknown Insurance Coverage, No Insurance. These variables were highly correlated to each other as seen in the correlation

## **Correlation Heatmap**

Correlation Heatmap

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **No Mental Illness** | 0.04 | 0.72 | 0.07 | 881 |
| **Mental Illness** | 0.99 | 0.63 | 0.77 | 44414 |
| **Accuracy** |  |  | 0.63 | 45295 |
| **Macro Avg** | 0.51 | 0.67 | 0.42 | 45295 |
| **Weighted Avg** | 0.97 | 0.63 | 0.76 | 45295 |

Once all the relevant predictors were remaining, we built the following models to see how socioeconomic factors affect Mental Illness among Adults:

## **Logistic Regression Model (Balanced class weights)**

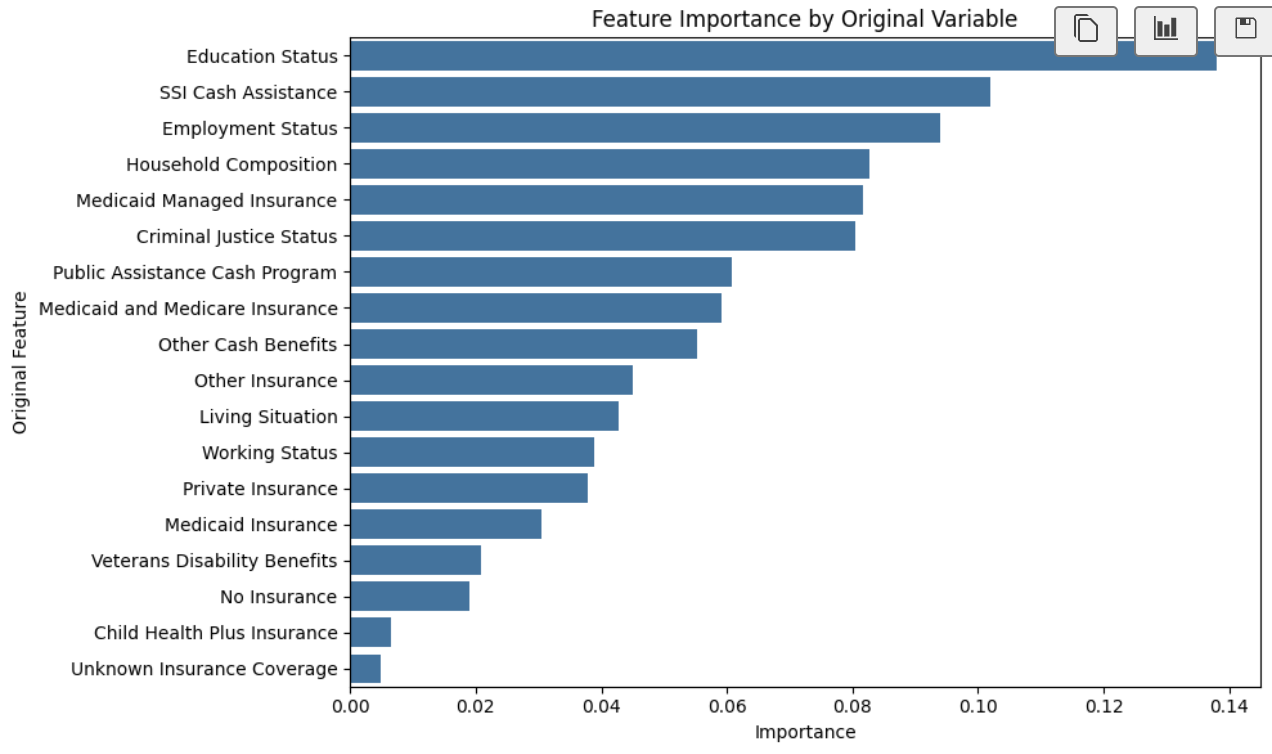
The balanced logistic regression model shows that it performs well in identifying the majority class (Mental Illness), achieving very high precision (0.99) and a strong recall (0.63) for this class. This means when the model predicts “1”, it is almost always correct, and it captures most of the actual 1’s. However, performance on the minority class (No Mental Illness) remains weak: precision is extremely low (0.04), meaning most predicted 0’s are actually incorrect, and although recall is higher (0.72), the very low precision results in a poor F1-score (0.07). Overall accuracy is 0.63, but this is driven almost entirely by the model’s success in the majority class. These results indicate that even with class balancing, logistic regression struggles to meaningfully learn the minority “0” class due to severe class imbalance and limited separation in the features. To improve this, we tried using the Random Forest Model again with balanced class weights.

## **Random Forest Model (Balanced class weights)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **No Mental Illness** | 0.04 | 0.57 | 0.07 | 881 |
| **Mental Illness** | 0.99 | 0.72 | 0.83 | 44414 |
| **Accuracy** |  |  | 0.72 | 45295 |
| **Macro Avg** | 0.51 | 0.65 | 0.45 | 45295 |
| **Weighted Avg** | 0.97 | 0.72 | 0.82 | 45295 |

The class-weighted Random Forest shows some improvement over logistic regression, especially in identifying the majority class. Precision for class 1 remains extremely high (0.99), and recall increases to 0.72, leading to a strong F1-score of 0.83. Performance on the minority class is still limited due to the very small support, recall changes from 0.72 (logistic) to 0.58, while precision remains low (0.04). Overall accuracy rises to 0.72, reflecting better learning of nonlinear patterns in the data but again for majority class only. Although the minority class is still challenging, Random Forest provides a better performance than logistic regression under severe class imbalance.

## **Feature importances for Random Forest Model (Balanced class weights)**



The Random Forest shows that several socioeconomic factors contribute to predicting mental health diagnoses. Education status is the strongest overall predictor, followed closely by Medicaid-related insurance and SSI cash assistance, indicating that financial and support-related factors matter most. Criminal justice status, living situation, and working status also provide moderate signals, while other benefit programs and child health insurance contribute smaller effects. Overall, the model uses many factors, but each provides only modest predictive power.

***NOTE:*** This feature importance only visualizes what features the current Random Forest Model is using the most to predict Mental Illness under class imbalance. These features are not necessarily the causal drivers.

## **Smart question 2 conclusion**

Overall, predicting mental illness using the available socioeconomic factors is challenging due to the strong class imbalance in the data. Logistic regression has difficulty identifying the minority class, and while the Random Forest performs better, it still cannot consistently detect class 0. The feature importance results indicate that Education status, Medicaid Insurance, and SSI Cash Assistance are the most influential predictors, with the remaining socioeconomic factors contributing smaller effects. The models capture some meaningful patterns, but the predictors offer only limited power for accurate classification. This suggests that socioeconomic factors, especially education attainment, can provide some insight into mental health status among adults in New York, but the class imbalance must be considered. For more reliable results, a balanced dataset would be necessary.

# **SMART Question 3**

How effectively can employment and related socioeconomic variables predict an individual’s living situation category in New York State using PCS 2019?

*Predicting variables:*

Employment Status, Age Group, Education Status, "SSI Cash Assistance, SSDI Cash Assistance, Public Assistance Cash Program, Other Cash Benefits, Medicaid Insurance, Medicare Insurance, Private Insurance

*Target Variable:* Living Situation

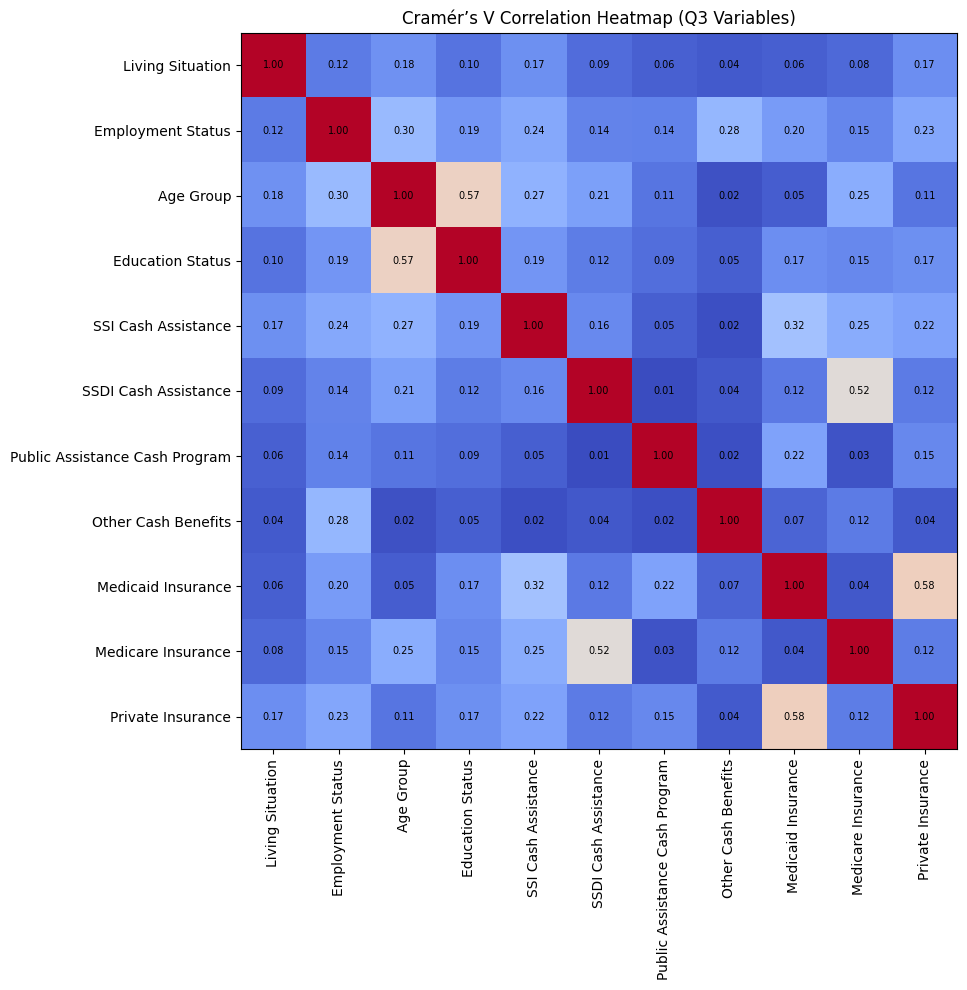
## **Data Description**

|  |  |
| --- | --- |
| **Living Situation** | |
| **PRIVATE RESIDENCE** | 107357 |
| **OTHER LIVING SITUATION** | 24162 |
| **INSTITUTIONAL SETTING** | 980 |

Living Situation is highly imbalanced, with Private Residence taking up most of the data and Institutional Setting appearing in less than 1%, which makes models lean toward the majority class. Later we apply class weighting and SMOTE to ease this issue.

## **Correlation test: Cramér’s V matrix**

We can observe the correlations between the variables from this chart.



None of the predictors show high correlation in the Cramér’s V matrix, and almost all values stay well below 0.60. This indicates that there is no strong multicollinearity issue among the variables.

In addition, our analysis does not use linear regression models, so multicollinearity would not be a major concern in the first place. Based on these results, all variables were kept for modeling.

After completing the necessary tests, we found that all of our variables are categorical. Therefore, we use the following models for our analysis:

## **Multinomial Logistic Regression**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** |
| **INSTITUTIONAL SETTING** | 0.02 | 0.55 | 0.04 |
| **OTHER LIVING SITUATION** | 0.27 | 0.57 | 0.37 |
| **PRIVATE RESIDENCE** | 0.93 | 0.48 | 0.64 |
| **Accuracy** |  |  | 0.5 |
| **Macro avg** | 0.41 | 0.54 | 0.35 |
| **Weighted avg** | 0.8 | 0.5 | 0.58 |

The logistic regression model does not perform well, with an overall accuracy of about 0.49. The model predicts the majority class (Private Residence) with high precision (0.93), but still has low recall (0.47), and the smaller classes perform much worse, especially Institutional, which has extremely low precision (0.02). Because the data is highly imbalanced and the groups overlap a lot, the next step is to try more flexible models and apply SMOTE to improve the minority-class predictions.

## **Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** |
| **INSTITUTIONAL SETTING** | 0.03 | 0.64 | 0.05 |
| **OTHER LIVING SITUATION** | 0.29 | 0.56 | 0.39 |
| **PRIVATE RESIDENCE** | 0.94 | 0.55 | 0.7 |
| **Accuracy** |  |  | 0.56 |
| **Macro avg** | 0.42 | 0.59 | 0.38 |
| **Weighted avg** | 0.82 | 0.56 | 0.64 |

The random forest model performed better than logistic regression, with an overall accuracy of about 0.56. For the largest group, Private Residence, precision stayed very high at 0.94, and recall was 0.55. The Other Living Situation group also improved, with a recall of 0.56. The Institutional Setting category remained the most difficult to predict. Its precision was still extremely low (0.03), but recall increased to 0.64, mainly because the number of real cases in the test set is very small. Overall, the random forest learned more structure than logistic regression, but the severe class imbalance still limited how well the model could predict the minority group.

## **Random Forest & SMOTE**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** |
| **INSTITUTIONAL SETTING** | 0.03 | 0.6 | 0.06 |
| **OTHER LIVING SITUATION** | 0.29 | 0.58 | 0.39 |
| **PRIVATE RESIDENCE** | 0.94 | 0.57 | 0.71 |
| **Accuracy** |  |  | 0.58 |
| **Macro avg** | 0.42 | 0.58 | 0.39 |
| **Weighted avg** | 0.81 | 0.58 | 0.65 |

After applying SMOTE to the training data, the random forest improved a bit. The overall accuracy went up to about 0.58. For the largest group, Private Residence, precision stayed high at 0.94 and recall increased to 0.58. The Other Living Situation group also performed better with a recall of 0.57. The Institutional Setting category still had very low precision (0.03), but its recall increased to 0.60, mainly because the test set contains very few actual cases from this group. The weighted F1-score was about 0.65, so the SMOTE + random forest model ended up being the most balanced option among the ones we tried.

## **Random Forest & SMOTE (Without minority class)**

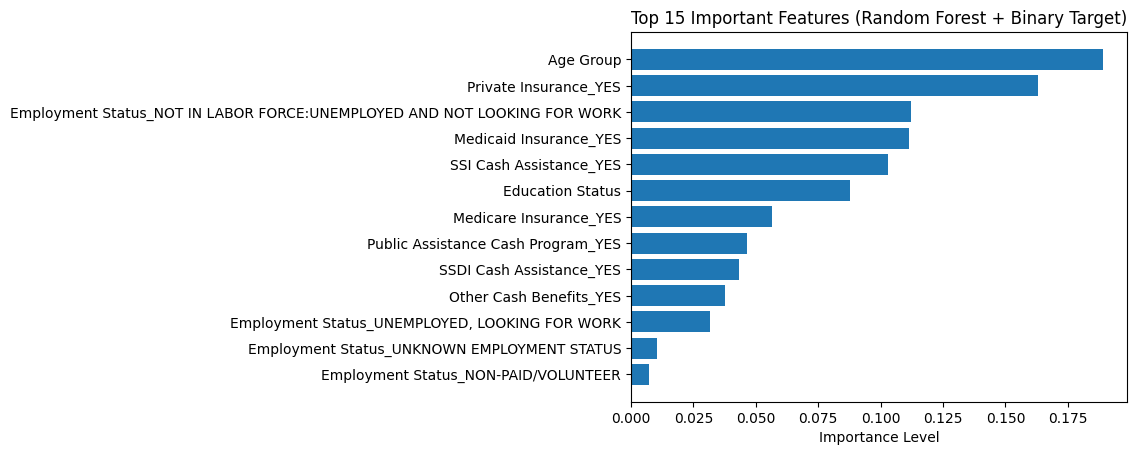
In the final step, we removed the minority categories in an effort to address the imbalance issue.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** |
| **OTHER LIVING SITUATION** | 0.33 | 0.8 | 0.47 |
| **PRIVATE RESIDENCE** | 0.93 | 0.64 | 0.76 |
| **Accuracy** |  |  | 0.67 |
| **Macro avg** | 0.63 | 0.72 | 0.61 |
| **Weighted avg** | 0.82 | 0.67 | 0.7 |

The model reached an accuracy of 0.67. For Private Residence, precision was high at 0.93 and recall was 0.64, so the model predicts this group fairly well. For Other Living Situation, recall was 0.80, but precision was lower at 0.33, meaning many predictions were incorrect. The weighted F1-score was 0.70, which shows the overall performance is acceptable after switching to a binary target. Compared to the earlier multi-class results, the binary version performs better, especially in recall.

## **Feature Importance for Random Forest & SMOTE (Without minority class)**

Since the Random Forest model with the removed minority classes and SMOTE produced the best results, we used this model to conduct the feature importance analysis.



Our results show that financial support and insurance coverage have a strong relationship with a person’s living situation. Age was the most important factor in our model, likely because people in different age groups face different levels of financial pressure and access to benefits. Insurance variables like private insurance, Medicaid, and Medicare were also high in importance. Cash assistance programs such as SSI and public assistance helped predict housing stability as well.

These results are similar to findings from earlier studies. Fenelon et al. (2017) reported that housing assistance helps people stay in stable living situations, and Friedman et al. (2022) found that people who rely on Medicaid often face more housing instability. Our model shows a similar pattern, where insurance and financial support play major roles in living conditions.

## **Smart question 3 conclusion**

Predicting the living situation in Q3 was difficult because the classes were very imbalanced and looked similar. Logistic regression didn’t work well, with an accuracy of about 0.49. Random forest did a little better, but still couldn’t identify the smallest group. After using SMOTE, the random forest improved to around 0.56 and became the most balanced model we had.

We also tried to decline the sample of larger groups to match the Institutional group, but that removed too much data and made the model unstable. So, the SMOTE random forest ended up being the most practical choice.

From the feature importance results, age was the strongest predictor. Insurance types like private insurance, Medicaid, and Medicare were also important, and cash assistance programs such as SSI and public assistance played a noticeable role. Overall, these factors helped the model separate the living situation categories better than the others.

# **Conclusion**

Across the three smart questions, the PCS dataset helped reveal clear patterns linking socioeconomic and demographic factors with both mental-illness reporting and living-situation categories. Although data imbalance limited prediction strength, the analyses consistently identified which variables had the greatest influence on each outcome.

For mental illness (Q1 & Q2), age group, education, household composition, employment status, and several insurance/assistance variables showed meaningful associations, indicating that social and economic stability is linked with how individuals report mental-health conditions.

For living situation (Q3), age group, insurance coverage, cash-assistance programs, and employment status were the strongest predictors, helping separate private residence, institutional settings, and other living arrangements.

Overall, the models highlighted the importance of insurance, financial support, age, and employment as the most consistent drivers across all three questions, providing useful insight into how socioeconomic conditions shape both mental-health outcomes and housing patterns in New York State.

# **Recommendations**

A main limitation in this project is the severe class imbalance across all three questions, which makes the smaller groups difficult for the models to learn and leads to unstable results. Another issue is that most variables in the PCS dataset are categorical and split into many small levels, increasing the chance of overfitting and limiting the models’ ability to capture stronger patterns.

For the next step, we could try more advanced imbalance-handling methods such as Balanced Random Forest or XGBoost with class weights to improve the performance for minority groups. Adding continuous variables, if they become available, may help strengthen the predictive signal. It may also be useful to analyze minority groups separately so their patterns are not overshadowed by the dominant majority class.

# **References**

Behavioral Health Barometer: United States, Volume 6. (n.d.). Substance Abuse and Mental Health Services Administration. <https://www.samhsa.gov/data/report/behavioral-health-barometer-united-states-volume-6>

Ettman, C. K., Cohen, G. H., & Galea, S. (2023). Financial strain and depression in the U.S.: A scoping review. Translational Psychiatry, 13(1). <https://pmc.ncbi.nlm.nih.gov/articles/PMC10182750/>

Fenelon, A., Slopen, N., Boudreaux, M., & Pollack, C. E. (2017). Housing assistance programs and adult health in the United States. American Journal of Public Health, 107(4), 571–578. <https://doi.org/10.2105/AJPH.2016.303649>

Friedman, C., Schiro, S., & Remmert, J. E. (2022). Housing insecurity of Medicaid beneficiaries with cognitive disabilities during the COVID-19 pandemic. Journal of Social Distress and the Homeless. Advance online publication. <https://doi.org/10.1080/10530789.2022.2096710>

Healthy People 2030. (n.d.). Housing instability. Office of Disease Prevention and Health Promotion. <https://health.gov/>

Office of the State of New York Comptroller. (2019). Housing affordability in New York State: June 2019. <https://www.osc.ny.gov/files/reports/special-topics/pdf/housing-affordability-2019.pdf>

Katon, W. J. (2011). Epidemiology and treatment of depression in patients with chronic medical illness. Dialogues in Clinical Neuroscience, 13(1), 7–23. <https://pubmed.ncbi.nlm.nih.gov/21485743/>

Maqbool, N., Viveiros, J., & Ault, M. (2015). The impacts of affordable housing on health: A research summary. National Housing Conference.

Mason, K. E., Baker, E., Blakely, T., & Bentley, R. (2013). Housing affordability and mental health: Does the relationship differ for renters and home purchasers? Social Science & Medicine, 94, 91–97. <https://pubmed.ncbi.nlm.nih.gov/23931950/>

Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Üstün, B. (2007). Depression, chronic diseases, and decrements in health: Results from the World Health Surveys. The Lancet, 370(9590), 851–858. <https://pubmed.ncbi.nlm.nih.gov/17826170/>

Tanarsuwongkul, S., et al. (2025). Associations between social determinants of health and mental health disorders among U.S. adults. Epidemiology and Psychiatric Sciences.