Mental Illness Prediction

Team Members:

Eshwar Reddy Boojanoor - boojanoor.eshwar-reddy@edu.dsti.institute

Kiran Kumar Thunga – thunga.kiran-kumar@edu.dsti.institute

Raghavendra Krishna Sai Kagitha - raghavendra-krishna-sai.kagitha@edu.dsti.institute

GitHub Link

Eshwar Reddy Boojanoor - https://github.com/eshwarreddy9948-hub/Mental-Ilness-prediction-DSTI-

Kiran Kumar Thunga – https://github.com/kirantunga1512-tech/Mental-illness-prediction-DSTI-

Raghavendra Krishna Sai Kagitha - https://github.com/kagitharaghavendra/Mental-illness-prediction-DSTI-

Project Context

- Mental health has become one of the most critical public health challenges in recent times.
- The New York State Office of Mental Health (NYS OMH) is the main public agency in New York that plans, funds, and provides mental health services across the state.
- Its mission is to promote mental health and recovery for people of all ages—especially adults with serious mental illness and children with severe emotional disturbances

- To understand the needs of people using the public mental health system, the NYS OMH runs the Patient Characteristics Survey (PCS) every two years.
- This survey gives a one-week snapshot of all patients who received mental health services.
- In 2019, the PCS collected data from around 196,000 clients across 4,000 programs, which is roughly 1% of the state's population

- The dataset we are working on comes from this survey. It captures not just clinical details (like whether a person has a mental illness diagnosis) but also social, demographic, and economic factors such as: Age, sex, race, ethnicity, and language preferences, Living situation, employment, and education, Co-occurring conditions like diabetes, substance use, or disabilities Insurance and support programs

Problem Statement

- The goal of this project is to predict whether a patient has a mental illness based on the information provided in the dataset.
- The dataset has 76 features (columns) and 196,102 rows of patient records
- The target variable is the “Mental Illness” column, which indicates whether the person is identified as having a mental illness.
- This problem is framed as a binary classification task:
- Yes → Patient has a mental illness
- No → Patient does not have a mental illness

Why It Matters

- Building such a model is not meant to replace clinical judgment but to support policymakers, researchers, and care providersby identifying patterns in the data.

For example:

Which groups of people are more likely to have mental illness?

How do factors like age, region, insurance coverage, or chronic medical conditions influence outcomes?

Can predictive models help in planning resources and services better?

Data Pipeline: Cleaning, preprocessing, and exploratory analysis.

There are 76 columns in dataset. We went through the data dictionary and found that most of the columns have YES/NO/UNKNOWN values.

However, there are some columns which have more categories, if they are selected in feature engineering, encoding them will increase the cardinality for model training, risking overfitting of model and impacting model performance and execution time.

We looked into those columns and tried grouping the categories meaningfully so that data quality is maintained.

Also, we need to check class imbalance in each of the column as well - before and after modifications, so that we can group reasonably to solve the prediction problem better.

There are some irrelevant columns like zip code and survey year. We don’t need them for making predictions, so we will drop them.

In the target variable – **“Mental Illness**” 99% of data falls under – YES/NO categories and only 1% of data is UNKNOWN.

We have removed the UNKNOWN values, to make this a binary classification problem.

All these 22 columns that have more than 3 categories - we will check them further and group them meaningfully to reduce cardinality

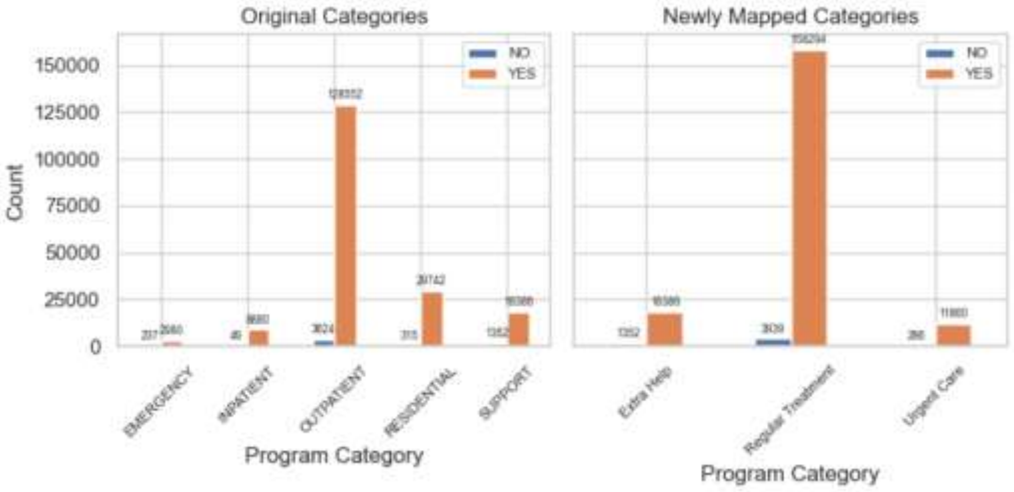
"Program Category", "Age Group", "Sex", "Transgender", "Region Served", "Sexual Orientation", "Hispanic Ethnicity", "Race", "Living Situation","Household Composition", "Preferred Language", "Religious Preference", "Special Education Services", "Mental Illness", "Employment Status","Number Of Hours Worked Each Week", "Education Status", "Unknown Chronic Med Condition", "Principal Diagnosis Class", "Medicaid Managed Insurance", "Additional Diagnosis Class", "Unknown Insurance Coverage"

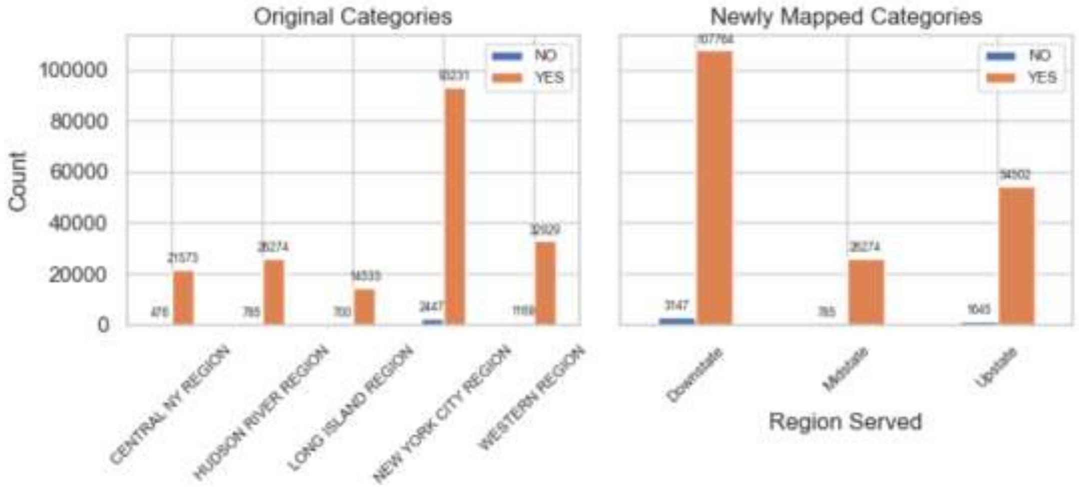
ALL COLUMNS THAT HAVE ENTRIES - YES, NO and UNKNOWN only (3 categories) - 52 columns

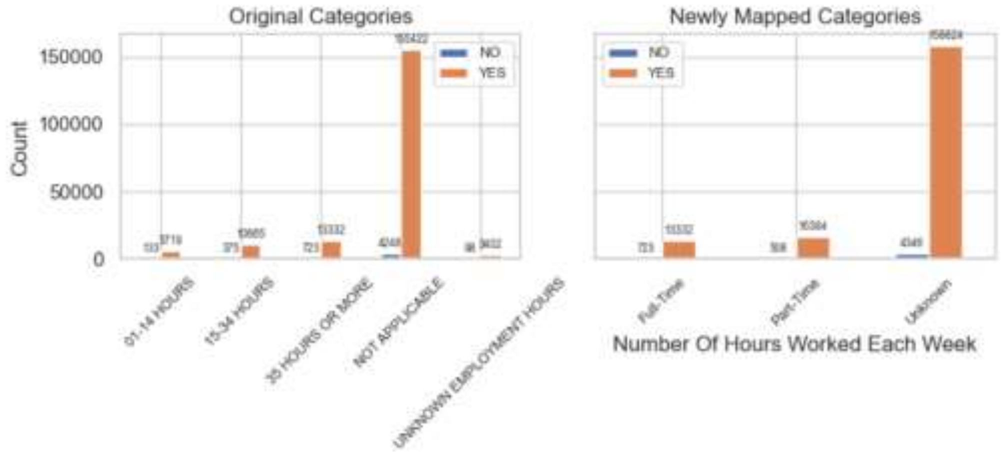
- We will group them and seperately check further EDAs on them.

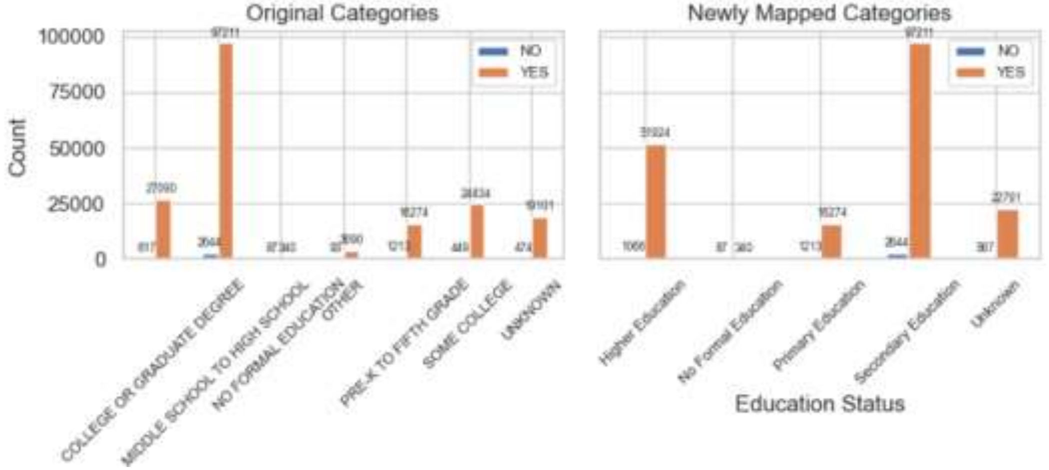
"Opioid Related Disorder", "Mobility Impairment Disorder", "Hearing Impairment", "Visual Impairment", "Speech Impairment", "Hyperlipidemia", "High Blood Pressure", "Diabetes", "Obesity", "Intellectual Disability", "Autism Spectrum", "Other Developmental Disability", "Alcohol Related Disorder", "Drug Substance Disorder", "Stroke", "Other Cardiac", "Pulmonary Asthma", "Alzheimer or Dementia", "Kidney Disease", "Liver Disease", "Endocrine Condition", "Neurological Condition", "Traumatic Brain Injury", "Joint Disease", "Cancer", "Heart Attack", "No Chronic Med Condition", "Other Chronic Med Condition", "Cannabis Recreational Use", "Cannabis Medicinal Use", "Smokes", "Received Smoking Medication", "Received Smoking Counseling", "Serious Mental Illness","Alcohol 12m Service", "Opioid 12m Service", "Drug/Substance 12m Service", "SSI Cash Assistance", "SSDI Cash Assistance", "Veterans Disability Benefits", "Veteran Status", "Veterans Cash Assistance", "Public Assistance Cash Program", "Other Cash Benefits", "Medicaid and Medicare Insurance", "No Insurance", "Medicaid Insurance","Medicare Insurance", "Private Insurance", "Child Health Plus Insurance", "Other Insurance", "Criminal Justice Status“

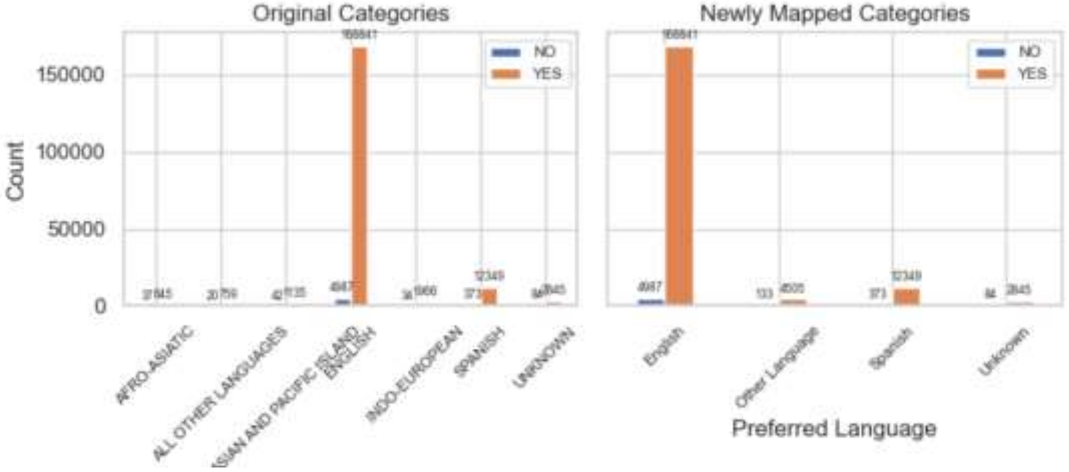
- All the columns that had more than 3 categories are first converted into bins/groups so that the imbalance in each field can be addressed.
- These are the before and after conversion class imbalance checks across various fields

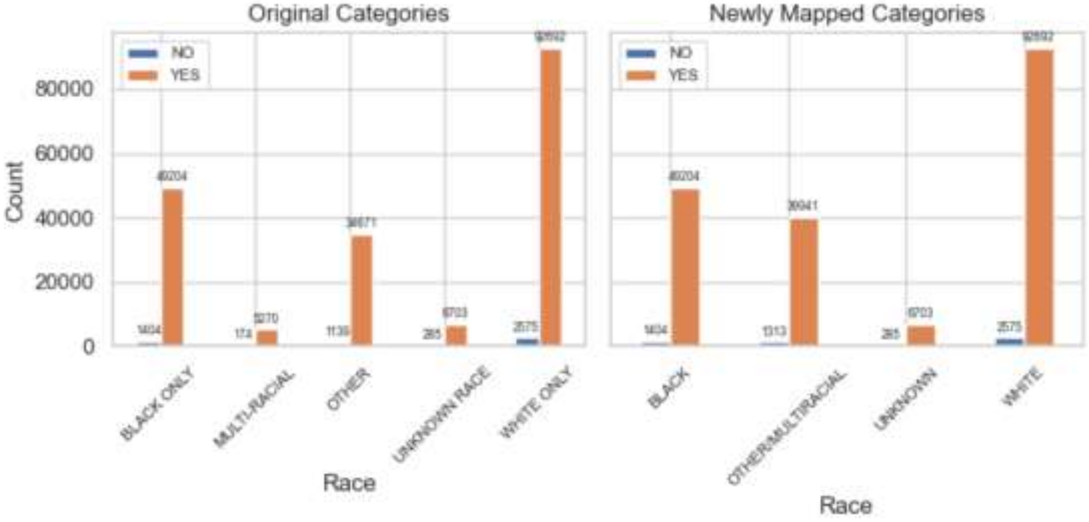


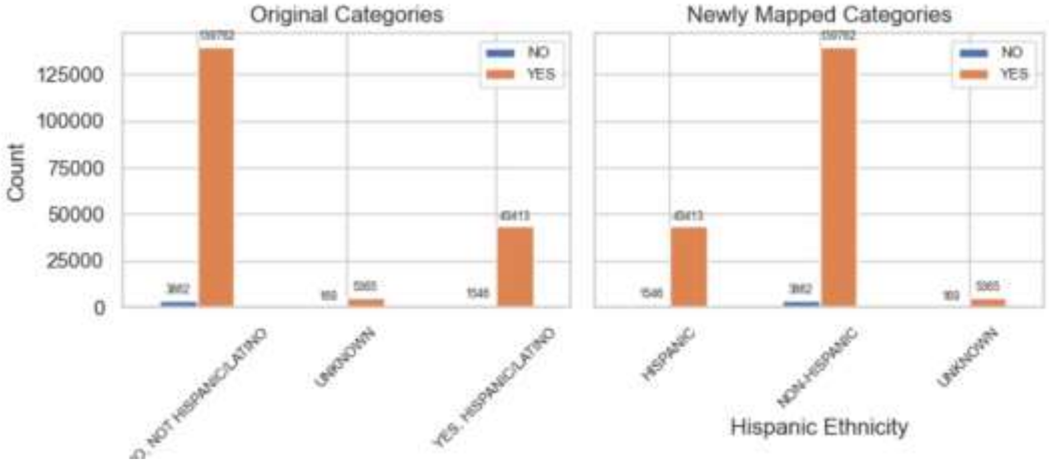


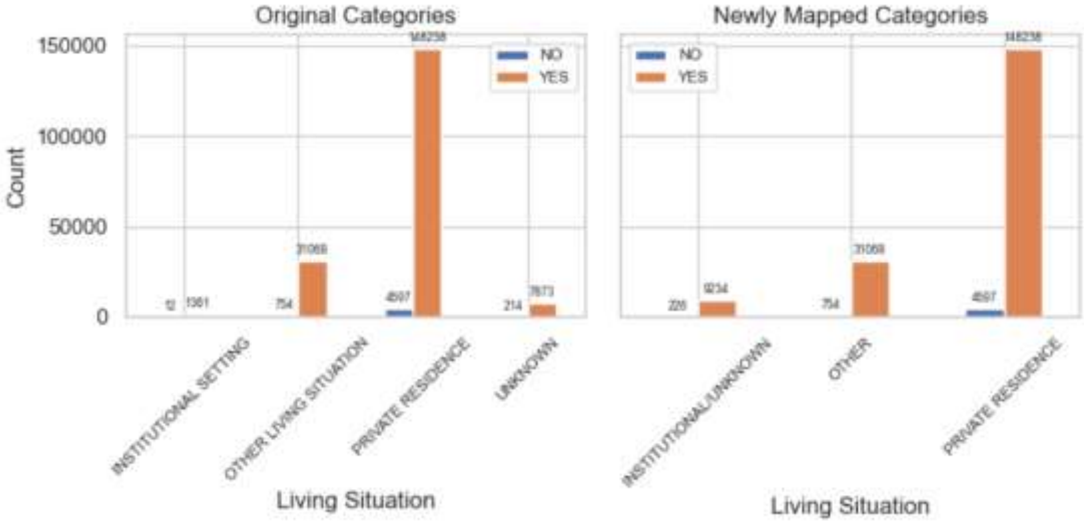


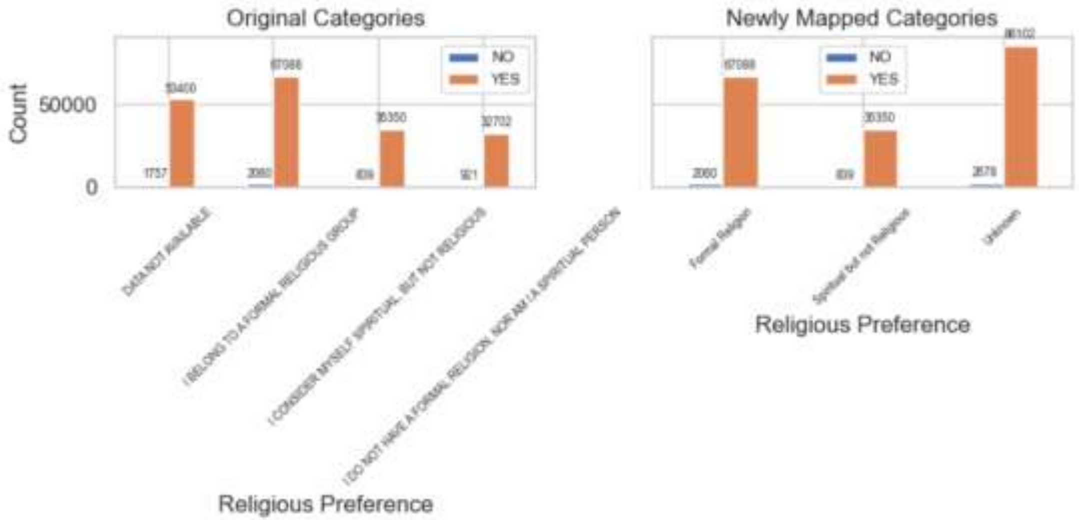


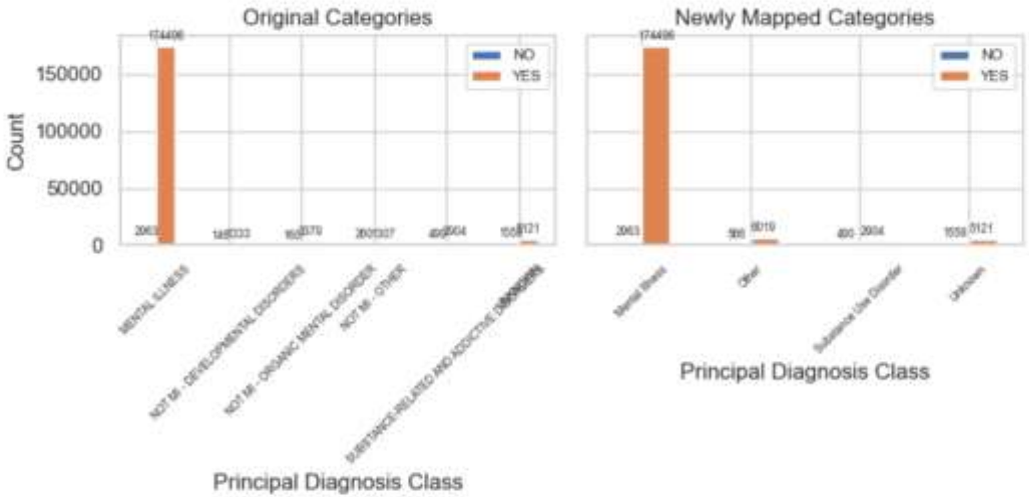

Original Categories
Newly Mapped Categories
Count
150000
100000
50000
0
NO
YES
37
845
20
759
42
1135
4987
168841
38
1966
373
12349
84
2845
AFRO-ASIATIC
ALL OTHER LANGUAGES
ASIAN AND PACIFIC ISLAND
ENGLISH
INDO-EUROPEAN
SPANISH
UNKNOWN
4987
168841
133
4505
373
12349
84
2845
English
Other Language
Spanish
Unknown
Preferred Language

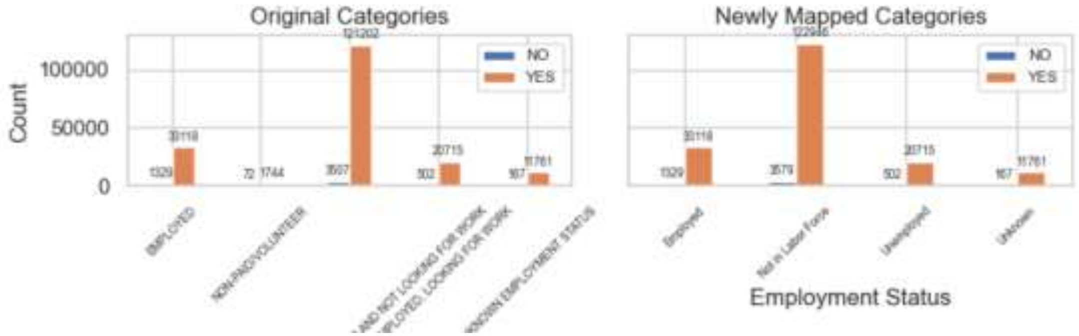

Original Categories
Newly Mapped Categories
Count
80000
60000
40000
20000
0
NO
YES
1404
49204
174
5270
1139
34671
285
6703
2575
92692
BLACK ONLY
MULTI-RACIAL
OTHER
UNKNOWN RACE
WHITE ONLY
Race
1404
49204
1313
39941
285
6703
2575
92692
BLACK
OTHER/MULTIRACIAL
UNKNOWN
WHITE
Race

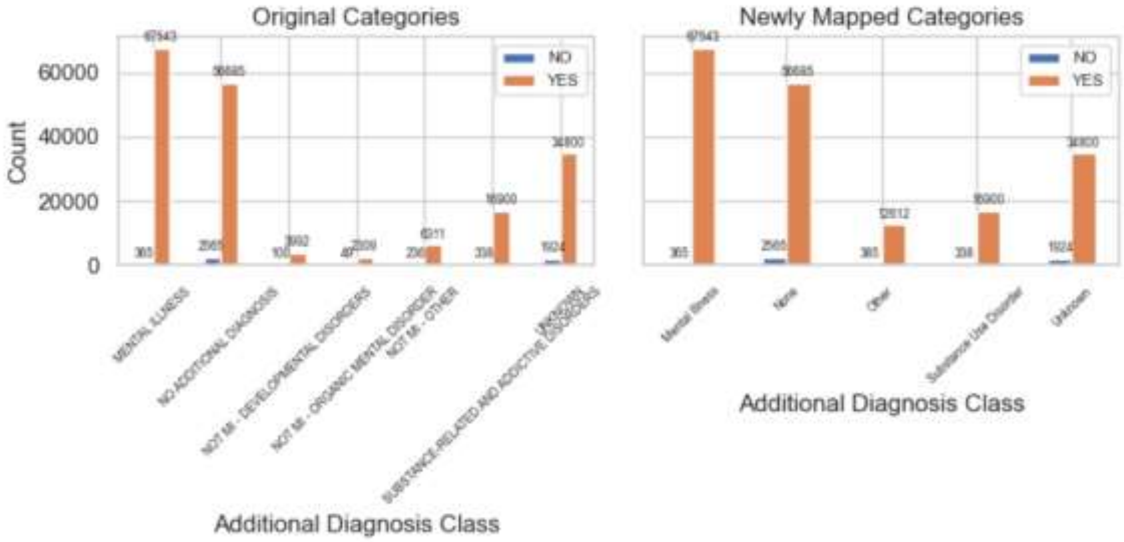

Original Categories
Newly Mapped Categories
Count
125000
100000
75000
50000
25000
0
NO
YES
3862
139762
169
5365
1546
43413
NO, NOT HISPANIC/LATINO
UNKNOWN
YES, HISPANIC/LATINO
1546
43413
3862
139762
169
5365
HISPANIC
NON-HISPANIC
UNKNOWN
Hispanic Ethnicity


Original Categories
Newly Mapped Categories
Count
150000
100000
50000
0
NO
YES
12
1361
754
31069
4597
148238
214
7873
INSTITUTIONAL SETTING
OTHER LIVING SITUATION
PRIVATE RESIDENCE
UNKNOWN
Living Situation
226
9234
754
31069
4597
148238
INSTITUTIONAL/UNKNOWN
OTHER
PRIVATE RESIDENCE
Living Situation


Original Categories
Newly Mapped Categories
Count
50000
0
NO
YES
1757
53400
2060
67088
839
35350
921
32702
DATA NOT AVAILABLE
I BELONG TO A FORMAL RELIGIOUS GROUP
I CONSIDER MYSELF SPIRITUAL, BUT NOT RELIGIOUS
I DO NOT HAVE A FORMAL RELIGION, NOR AM I A SPIRITUAL PERSON
Religious Preference
2060
67088
839
35350
2678
86102
Formal Religion
Spiritual but not Religious
Unknown
Religious Preference


Original Categories
Newly Mapped Categories
Count
150000
100000
50000
0
NO
YES
2063
174496
490
2904
1559
3121
MENTAL ILLNESS
NOT MI - DEVELOPMENTAL DISORDERS
NOT MI - ORGANIC MENTAL DISORDER
NOT MI - OTHER
SUBSTANCE-RELATED AND ADDICTIVE DISORDERS
UNKNOWN
Principal Diagnosis Class
2063
174496
586
8019
490
2904
1559
3121
Mental Illness
Other
Substance Use Disorder
Unknown
Principal Diagnosis Class


Original Categories
Newly Mapped Categories
Count
100000
50000
0
NO
YES
1329
33118
72
1744
3507
121202
502
20715
167
11761
EMPLOYED
NON-PAID/VOLUNTEER
AND NOT LOOKING FOR WORK
EMPLOYED, LOOKING FOR WORK
KNOWN EMPLOYMENT STATUS
1329
33118
3579
122946
502
20715
167
11761
Employed
Not in Labor Force
Unemployed
Unknown
Employment Status


Original Categories
Newly Mapped Categories
Count
60000
40000
20000
0
NO
YES
365
67543
2565
56685
100
3992
47
2509
236
6211
338
16900
1924
34800
MENTAL ILLNESS
NO ADDITIONAL DIAGNOSIS
NOT MI - DEVELOPMENTAL DISORDERS
NOT MI - ORGANIC MENTAL DISORDER
NOT MI - OTHER
SUBSTANCE-RELATED AND ADDICTIVE DISORDERS
UNKNOWN
Additional Diagnosis Class
365
67543
2565
56685
385
12512
338
16900
1924
34800
Mental Illness
None
Other
Substance Use Disorder
Unknown
Additional Diagnosis Class

Feature Engineering Summary

- Few columns are merged together to create a final column that captures the summary of those columns and dropped the original columns to reduce cardinality.
- Many columns that have 3 categories are grouped in this manner so that overall columns and categories in each columns can be summarised and grouped as much as possible.
- After all this process, 76 columns are finally down to 38 columns in EDA and feature engineering process

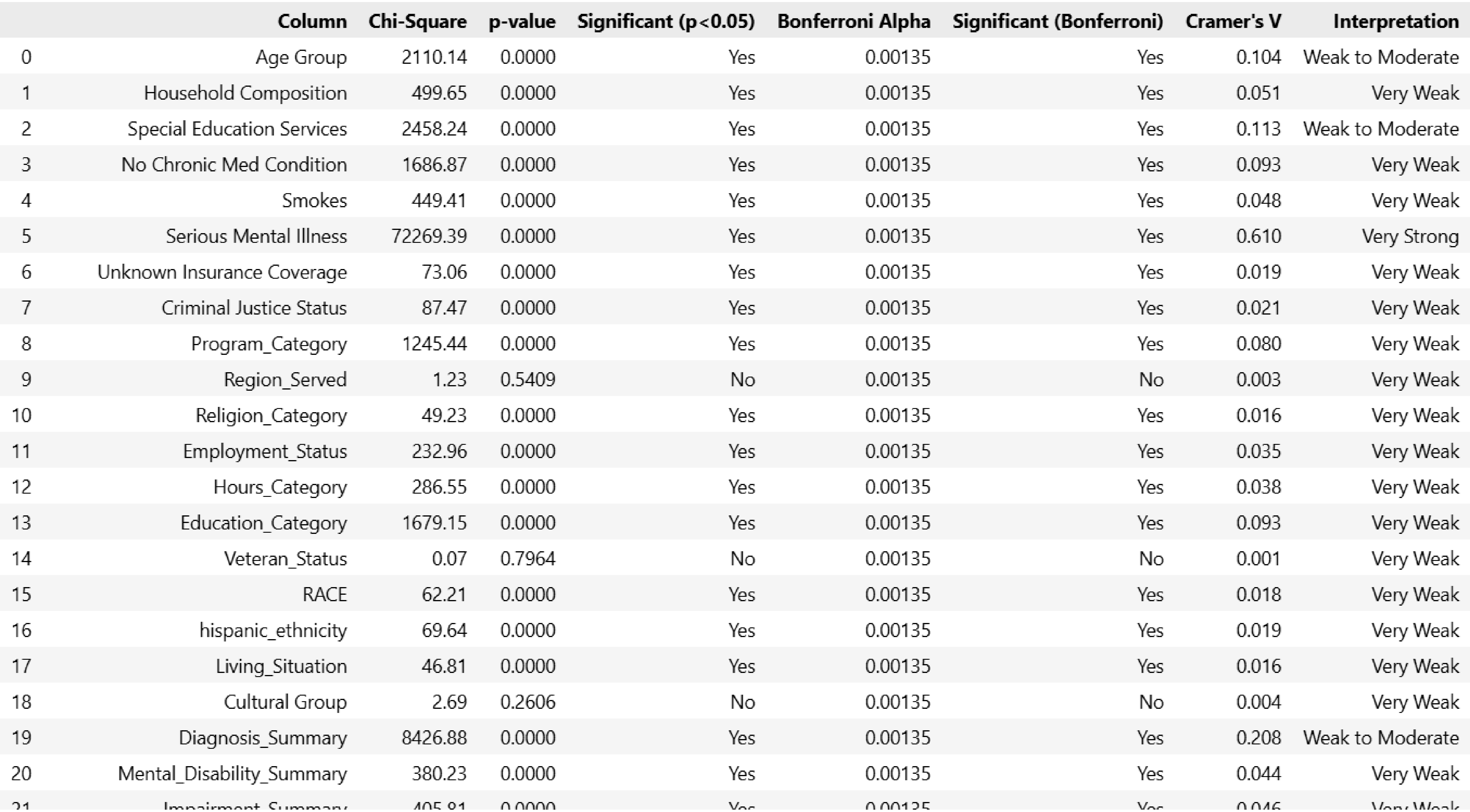
Feature Selection

- Statistical tests are conducted on each of the columns and checked the significance. There are 3 non significant columns, those are removed – Chi Square test, Bonferroni correction and creamer's V methods

- During Chi-Square feature selection, the column "Serious Mental Illness“ showed a very strong association with the target variable (Cramer's V = 0.61). While this indicates high predictive power, it also suggests strong information leakage, since "Serious Mental Illness" is essentially another way of stating the target outcome.

- To build a fair and generalizable model, we decided to DROP "Serious Mental Illness" from the modeling features. However, it is still reported in the EDA section to highlight its strong statistical significance.

- 'Cultural Group', 'Veteran_Status' , 'Region_Served' are the non-significant columns, dropped them and created final data for ML predictions



	Column	Chi-Square	p-value	Significant (p<0.05)	Bonferroni Alpha	Significant (Bonferroni)	Cramer's V	Interpretation
0	Age Group	2110.14	0.0000	Yes	0.00135	Yes	0.104	Weak to Moderate
1	Household Composition	499.65	0.0000	Yes	0.00135	Yes	0.051	Very Weak
2	Special Education Services	2458.24	0.0000	Yes	0.00135	Yes	0.113	Weak to Moderate
3	No Chronic Med Condition	1686.87	0.0000	Yes	0.00135	Yes	0.093	Very Weak
4	Smokes	449.41	0.0000	Yes	0.00135	Yes	0.048	Very Weak
5	Serious Mental Illness	72269.39	0.0000	Yes	0.00135	Yes	0.610	Very Strong
6	Unknown Insurance Coverage	73.06	0.0000	Yes	0.00135	Yes	0.019	Very Weak
7	Criminal Justice Status	87.47	0.0000	Yes	0.00135	Yes	0.021	Very Weak
8	Program_Category	1245.44	0.0000	Yes	0.00135	Yes	0.080	Very Weak
9	Region_Served	1.23	0.5409	No	0.00135	No	0.003	Very Weak
10	Religion_Category	49.23	0.0000	Yes	0.00135	Yes	0.016	Very Weak
11	Employment_Status	232.96	0.0000	Yes	0.00135	Yes	0.035	Very Weak
12	Hours_Category	286.55	0.0000	Yes	0.00135	Yes	0.038	Very Weak
13	Education_Category	1679.15	0.0000	Yes	0.00135	Yes	0.093	Very Weak
14	Veteran_Status	0.07	0.7964	No	0.00135	No	0.001	Very Weak
15	RACE	62.21	0.0000	Yes	0.00135	Yes	0.018	Very Weak
16	hispanic_ethnicity	69.64	0.0000	Yes	0.00135	Yes	0.019	Very Weak
17	Living_Situation	46.81	0.0000	Yes	0.00135	Yes	0.016	Very Weak
18	Cultural Group	2.69	0.2606	No	0.00135	No	0.004	Very Weak
19	Diagnosis_Summary	8426.88	0.0000	Yes	0.00135	Yes	0.208	Weak to Moderate
20	Mental_Disability_Summary	380.23	0.0000	Yes	0.00135	Yes	0.044	Very Weak
21	Impairment_Summary	405.81	0.0000	Yes	0.00135	Yes	0.046	Very Weak

Model Building

- Encoded all binary columns directly into 0,1
- Encoded Education status using Ordinal encoding, since order is needed.
- All other categorical columns with 3 or more are encoded using One Hot Encoding method.
- Final data size that was fed to model is around 91 columns.
- There is a huge class imbalance in dataset's Target variable – 97% is Yes and 3% is No. So all the initial iterations gave huge accuracy numbers which is not correct since model is biased towards positive class, making poor predictions on negative class.
- So, we used SMOTE method to handle this class imbalance. Models performed better after SMOTE technique.
- Tuned the models using GridSearchCV to figure out the best parameters other than baseline model parameters to increase model performance.
- All final models are exported after training to pickle files (.pkl) so that they can be called easily again for predictions.

Models Created

We built four models –

Three ML classification models

1. Logistic Regression – 1.5 hours
2. Random Forest - training time -25min
3. Decision Tree classifier – training time – 45min

A Nueral Network

MLPClassifier – Training time – 1hr 10min

Model Comparison Summary

Model	Stage	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Best Cross Validation F1 Scores
Random Forest	Baseline	97%	53%	14%	23%	98%	100%	99%	
Random Forest	After SMOTE	95%	21%	26%	24%	98%	97%	97%	
Random Forest	After Tuning	89%	15%	58%	24%	99%	90%	94%	93%
Decision Tree	Baseline	95%	20%	24%	22%	98%	97%	98%	
Decision Tree	After SMOTE	93%	16%	31%	21%	98%	95%	96%	
Decision Tree	After Tuning	91%	18%	36%	24%	98%	94%	96%	95%
Logistic Regression	Baseline	97%	60%	7%	13%	97%	100%	99%	
Logistic Regression	After SMOTE	94%	18%	34%	23%	98%	95%	97%	
Logistic Regression	After Tuning	95%	20%	40%	26%	98%	96%	97%	95%
Neural Network (MLP)	After SMOTE	95%	18%	26%	21%	98%	97%	97%	

Random Forest
Strong baseline accuracy (97%) but weak recall for the minority class (No Mental Illness).
After SMOTE, recall improved slightly for class 0 but overall performance dropped in tuning (accuracy 89%).
Best CV F1: 0.9337.

Decision Tree

Baseline accuracy of 95%, but also biased toward the majority class.
After SMOTE, recall for class 0 improved modestly.
After tuning, it achieved better balance, with Best CV F1: 0.9476.

Logistic Regression

High baseline accuracy (97%) but very poor recall for class 0.
After SMOTE, improved recall and more balanced results.
After tuning, it achieved the best overall performance with Best CV F1: 0.9515, showing stable precision and recall across both classes.

Neural Network (MLPClassifier)

After SMOTE, accuracy reached 95%.
It achieved an F1 score of 0.21 for class 0 and 0.97 for class 1.
While competitive with Decision Tree and Random Forest, it did not surpass Logistic Regression.

Best Model Choice

- Although Random Forest and Decision Tree performed reasonably well, they remained biased toward the majority class. The Neural Network gave balanced performance but did not outperform the simpler models.
- Logistic Regression emerged as the best model overall, achieving the highest cross-validated F1 score (0.9515) and offering the most consistent balance between precision and recall after tuning.
- If you consider training time as a constraint, Decision Tree classifier is the second best.

Web Application

- We used Flask framework to build a python based web application.

➢ It has two screens – Landing page and Form

➢ Clicking on Landing page takes you to second page

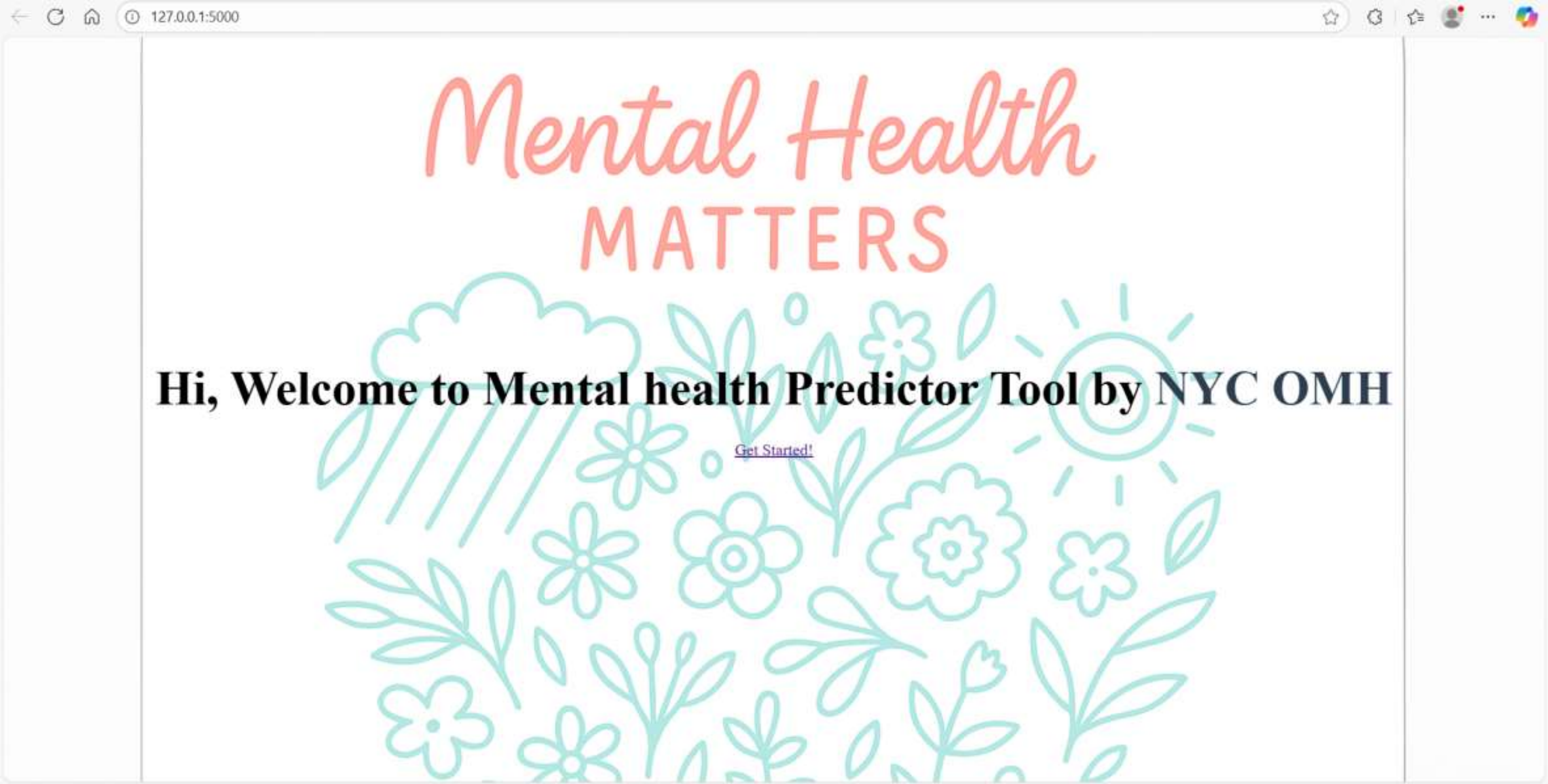
➢ In second page form, all the fields are mandatory to fill, they are the columns and categories we used in data that was there before data is encoded and fed to models.

➢ If any field is not filled, clicking on submit will prompt you to fill those details again.

➢ After all details are filled, the data will be collected

➢ This data can be sent to backend where we can have .pkl file predicting results directly and those results will have to be loaded again in a text box in App saying YES/NO for prediction.

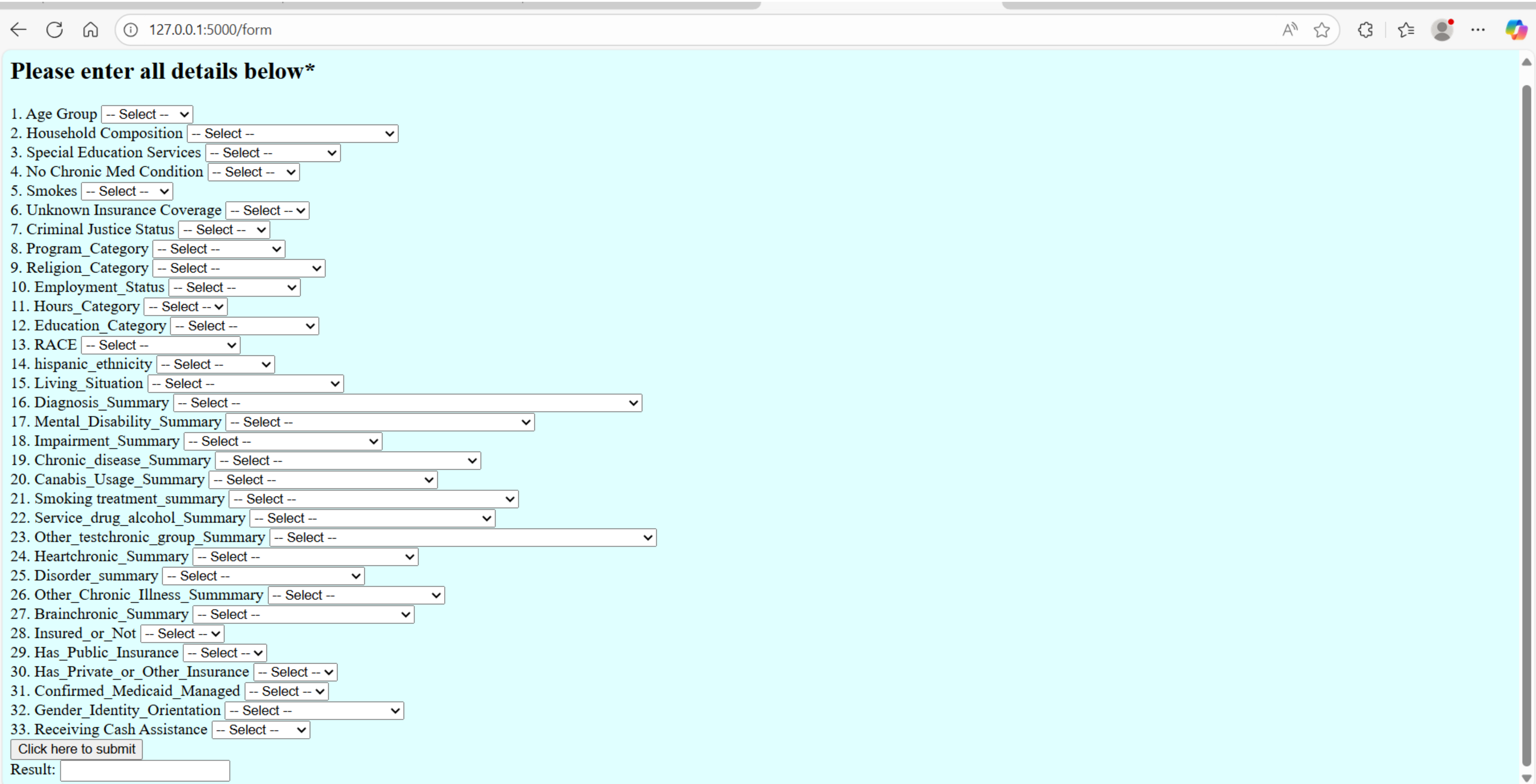
Landing Page



Hi, Welcome to Mental health Predictor Tool by NYC OMH

Get Started!

Page 2

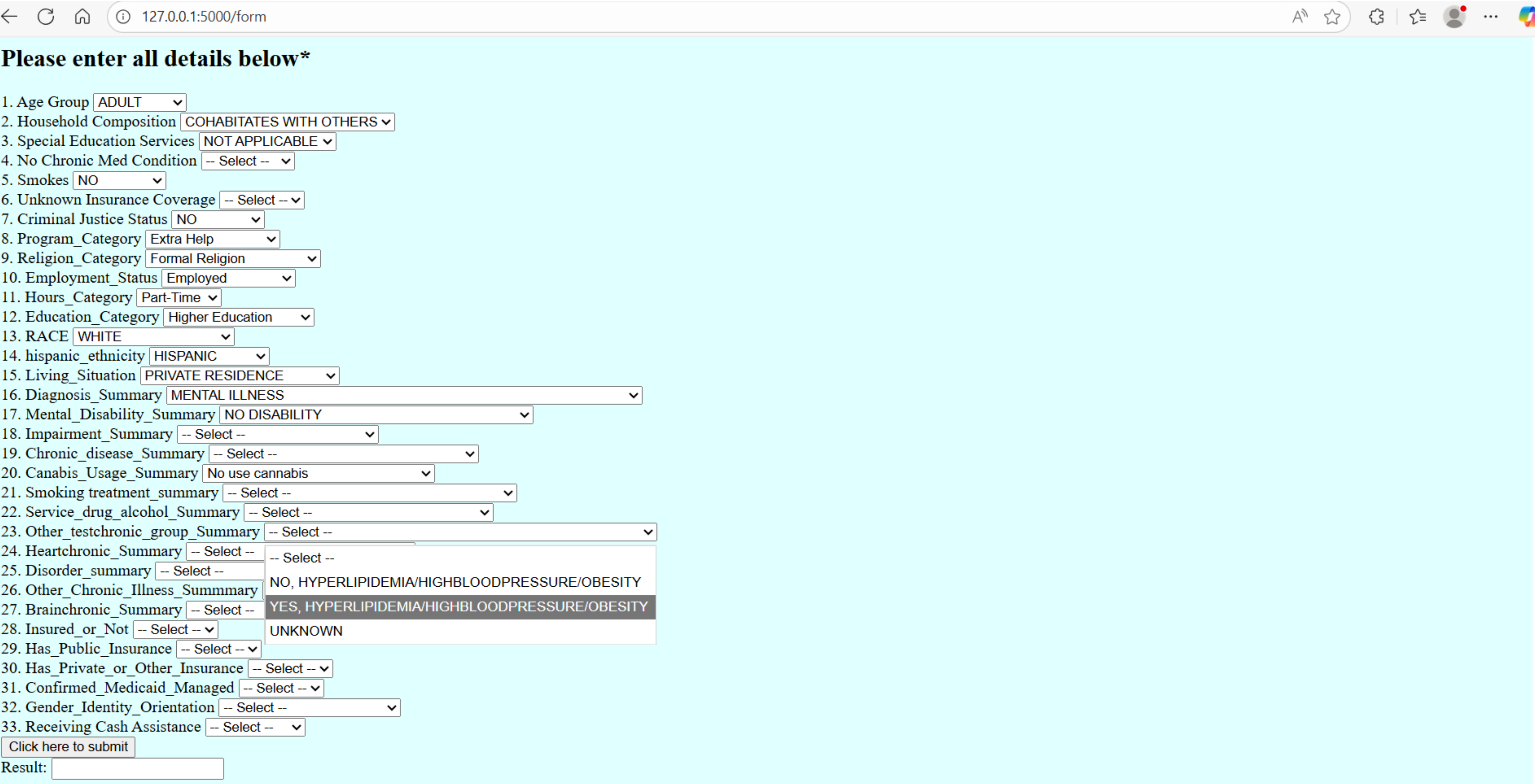


Please enter all details below*

1. Age Group -- Select --
2. Household Composition -- Select --
3. Special Education Services -- Select --
4. No Chronic Med Condition -- Select --
5. Smokes -- Select --
6. Unknown Insurance Coverage -- Select --
7. Criminal Justice Status -- Select --
8. Program_Category -- Select --
9. Religion_Category -- Select --
10. Employment_Status -- Select --
11. Hours_Category -- Select --
12. Education_Category -- Select --
13. RACE -- Select --
14. hispanic_ethnicity -- Select --
15. Living_Situation -- Select --
16. Diagnosis_Summary -- Select --
17. Mental_Disability_Summary -- Select --
18. Impairment_Summary -- Select --
19. Chronic_disease_Summary -- Select --
20. Canabis_Usage_Summary -- Select --
21. Smoking treatment_summary -- Select --
22. Service_drug_alcohol_Summary -- Select --
23. Other_testchronic_group_Summary -- Select --
24. Heartchronic_Summary -- Select --
25. Disorder_summary -- Select --
26. Other_Chronic_Illness_Summmary -- Select --
27. Brainchronic_Summary -- Select --
28. Insured_or_Not -- Select --
29. Has_Public_Insurance -- Select --
30. Has_Private_or_Other_Insurance -- Select --
31. Confirmed_Medicaid_Managed -- Select --
32. Gender_Identity_Orientation -- Select --
33. Receiving Cash Assistance -- Select --

Result:

Form filling with dropdown showing categories



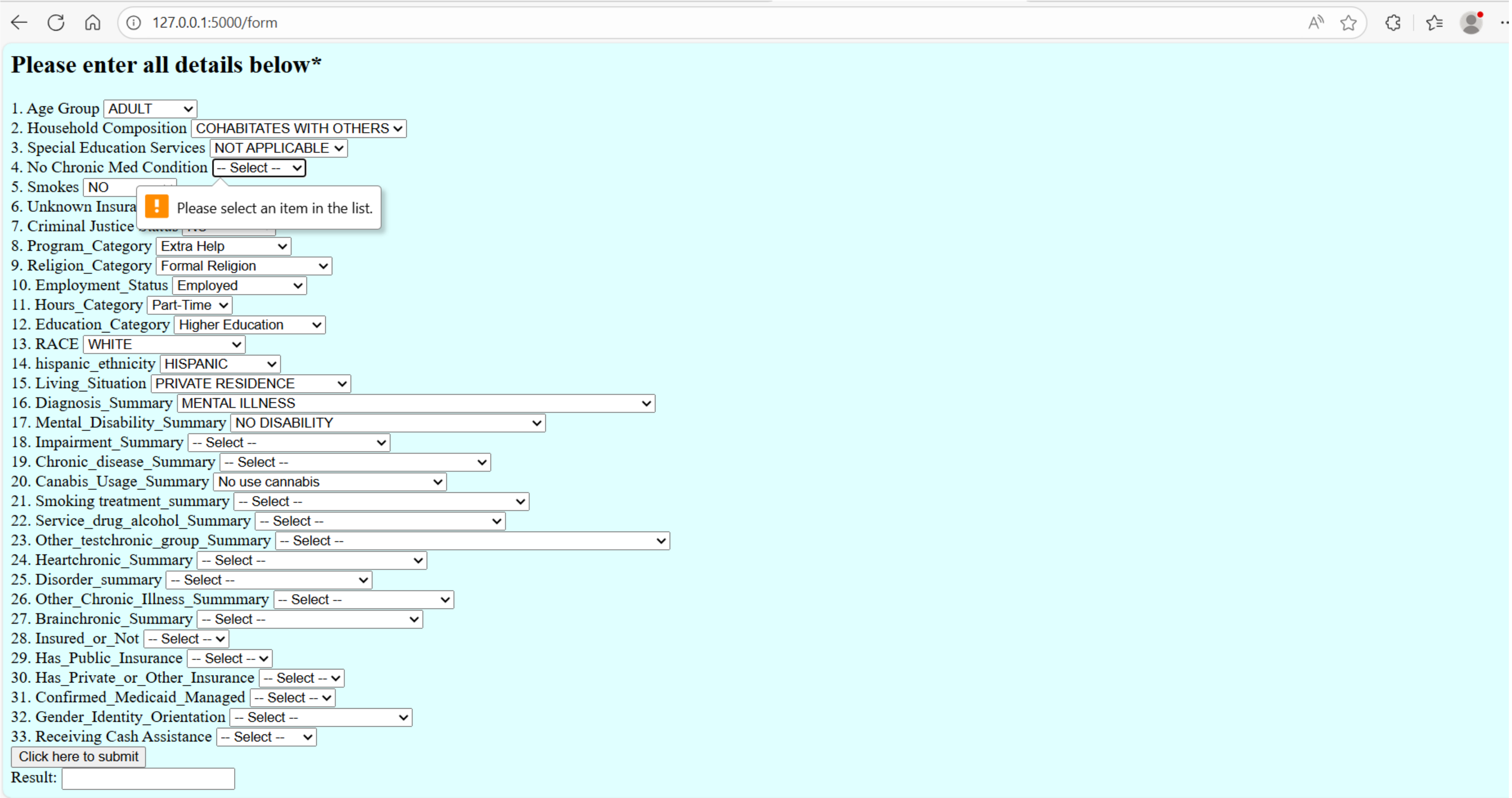
Please enter all details below*

1. Age Group ADULT
2. Household Composition COHABITATES WITH OTHERS
3. Special Education Services NOT APPLICABLE
4. No Chronic Med Condition -- Select --
5. Smokes NO
6. Unknown Insurance Coverage -- Select --
7. Criminal Justice Status NO
8. Program_Category Extra Help
9. Religion_Category Formal Religion
10. Employment_Status Employed
11. Hours_Category Part-Time
12. Education_Category Higher Education
13. RACE WHITE
14. hispanic_ethnicity HISPANIC
15. Living_Situation PRIVATE RESIDENCE
16. Diagnosis_Summary MENTAL ILLNESS
17. Mental_Disability_Summary NO DISABILITY
18. Impairment_Summary -- Select --
19. Chronic_disease_Summary -- Select --
20. Canabis_Usage_Summary No use cannabis
21. Smoking treatment_summary -- Select --
22. Service_drug_alcohol_Summary -- Select --
23. Other_testchronic_group_Summary -- Select --
24. Heartchronic_Summary -- Select --
25. Disorder_summary -- Select --
26. Other_Chronic_Illness_Summmary
27. Brainchronic_Summary -- Select --
28. Insured_or_Not -- Select --
29. Has_Public_Insurance -- Select --
30. Has_Private_or_Other_Insurance -- Select --
31. Confirmed_Medicaid_Managed -- Select --
32. Gender_Identity_Orientation -- Select --
33. Receiving Cash Assistance -- Select --

-- Select --
NO, HYPERLIPIDEMIA/HIGHBLOODPRESSURE/OBESITY
YES, HYPERLIPIDEMIA/HIGHBLOODPRESSURE/OBESITY
UNKNOWN
Click here to submit

Result:

Unfilled categories are warned again



Please enter all details below*

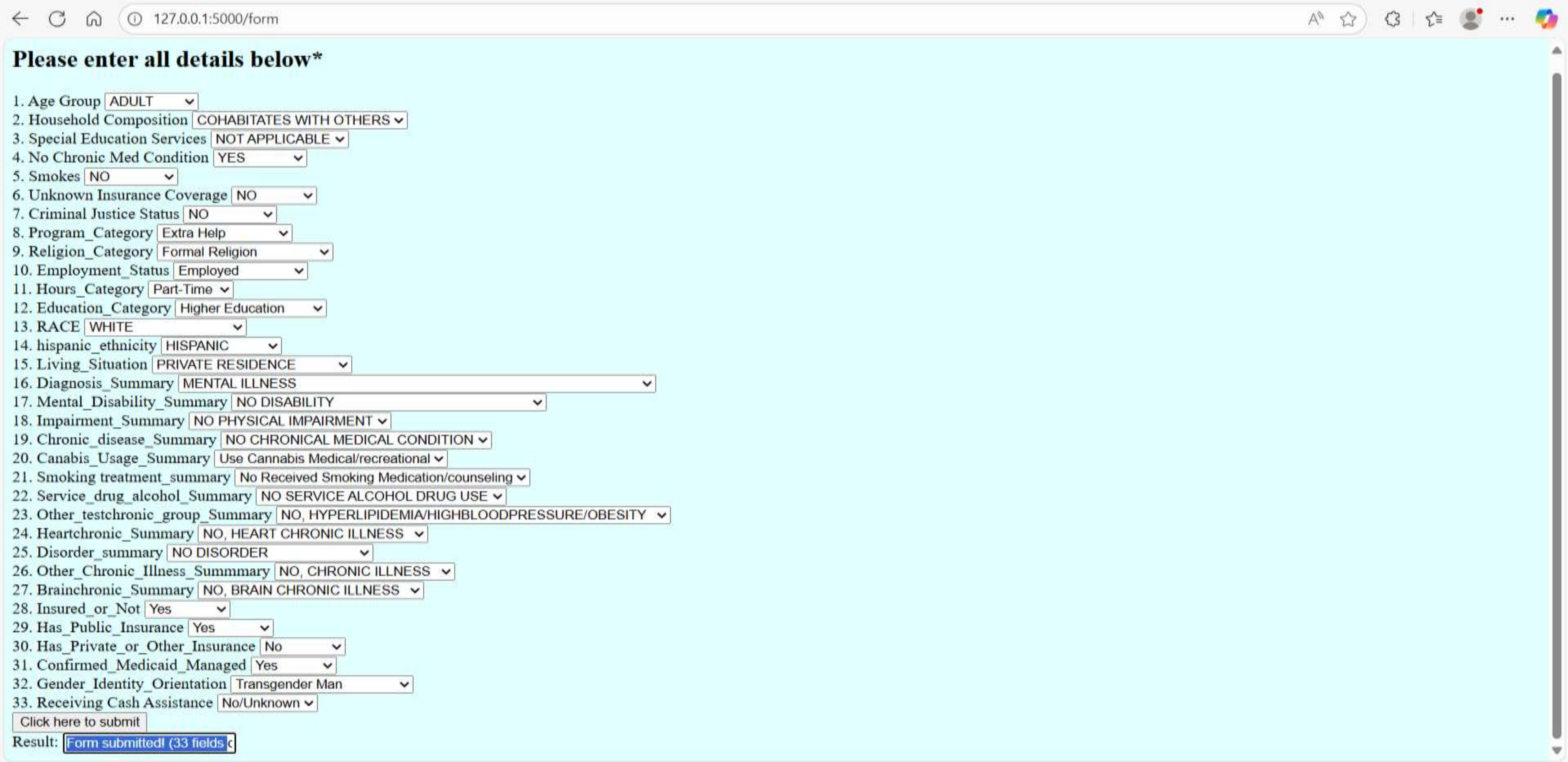
1. Age Group ADULT
2. Household Composition COHABITATES WITH OTHERS
3. Special Education Services NOT APPLICABLE
4. No Chronic Med Condition -- Select --
5. Smokes NO
6. Unknown Insura
7. Criminal Justice Status NO

Please select an item in the list.

8. Program_Category Extra Help
9. Religion_Category Formal Religion
10. Employment_Status Employed
11. Hours_Category Part-Time
12. Education_Category Higher Education
13. RACE WHITE
14. hispanic_ethnicity HISPANIC
15. Living_Situation PRIVATE RESIDENCE
16. Diagnosis_Summary MENTAL ILLNESS
17. Mental_Disability_Summary NO DISABILITY
18. Impairment_Summary -- Select --
19. Chronic_disease_Summary -- Select --
20. Canabis_Usage_Summary No use cannabis
21. Smoking treatment_summary -- Select --
22. Service_drug_alcohol_Summary -- Select --
23. Other_testchronic_group_Summary -- Select --
24. Heartchronic_Summary -- Select --
25. Disorder_summary -- Select --
26. Other_Chronic_Illness_Summmary -- Select --
27. Brainchronic_Summary -- Select --
28. Insured_or_Not -- Select --
29. Has_Public_Insurance -- Select --
30. Has_Private_or_Other_Insurance -- Select --
31. Confirmed_Medicaid_Managed -- Select --
32. Gender_Identity_Orientation -- Select --
33. Receiving Cash Assistance -- Select --

Result:

Form submission will give 33 columns collected at the end.



Please enter all details below*

1. Age Group ADULT
2. Household Composition COHABITATES WITH OTHERS
3. Special Education Services NOT APPLICABLE
4. No Chronic Med Condition YES
5. Smokes NO
6. Unknown Insurance Coverage NO
7. Criminal Justice Status NO
8. Program_Category Extra Help
9. Religion_Category Formal Religion
10. Employment_Status Employed
11. Hours_Category Part-Time
12. Education_Category Higher Education
13. RACE WHITE
14. hispanic_ethnicity HISPANIC
15. Living_Situation PRIVATE RESIDENCE
16. Diagnosis_Summary MENTAL ILLNESS
17. Mental_Disability_Summary NO DISABILITY
18. Impairment_Summary NO PHYSICAL IMPAIRMENT
19. Chronic_disease_Summary NO CHRONICAL MEDICAL CONDITION
20. Canabis_Usage_Summary Use Cannabis Medical/recreational
21. Smoking treatment_summary No Received Smoking Medication/counseling
22. Service_drug_alcohol_Summary NO SERVICE ALCOHOL DRUG USE
23. Other_testchronic_group_Summary NO, HYPERLIPIDEMIA/HIGHBLOODPRESSURE/OBESITY
24. Heartchronic_Summary NO, HEART CHRONIC ILLNESS
25. Disorder_summary NO DISORDER
26. Other_Chronic_Illness_Summmary NO, CHRONIC ILLNESS
27. Brainchronic_Summary NO, BRAIN CHRONIC ILLNESS
28. Insured_or_Not Yes
29. Has_Public_Insurance Yes
30. Has_Private_or_Other_Insurance No
31. Confirmed_Medicaid_Managed Yes
32. Gender_Identity_Orientation Transgender Man
33. Receiving Cash Assistance No/Unknown

Click here to submit

Result: Form submitted! (33 fields c