Susceptibility of Anxiety, Depression, and ADHD

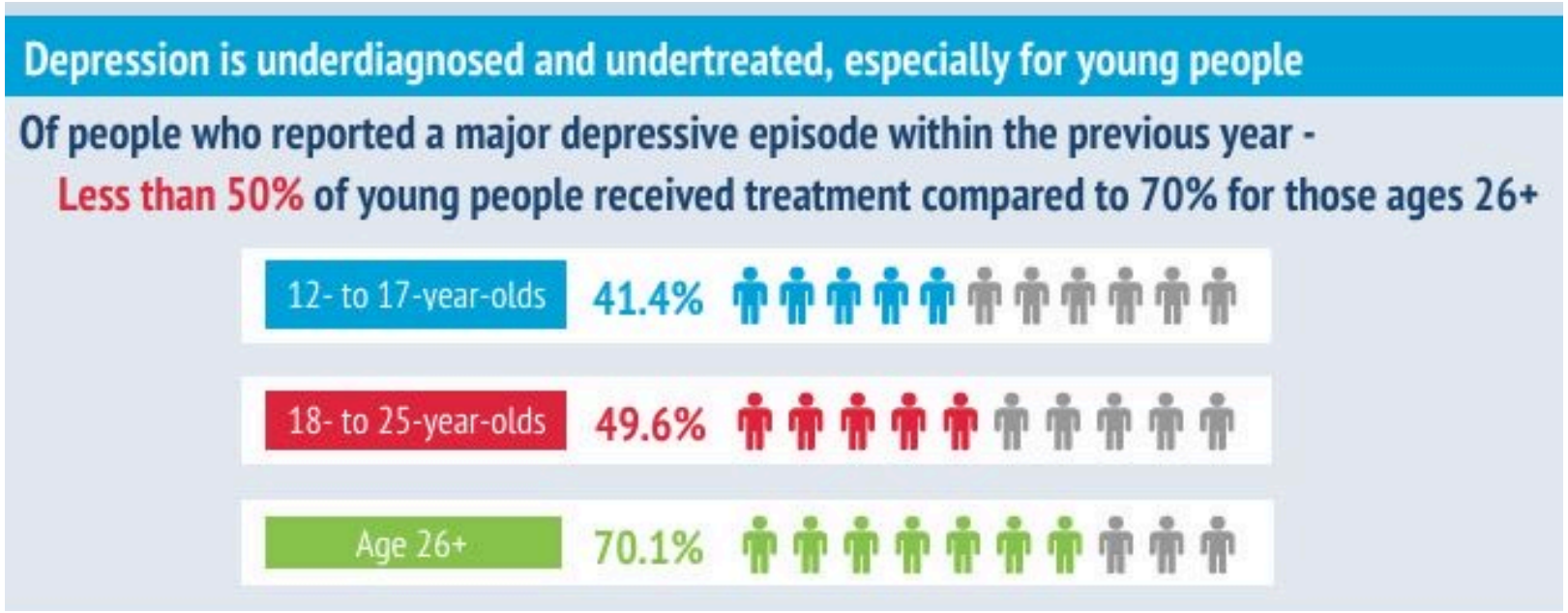
By: Shaili Gunda, Bill Lu, Mihir Padsumbiya, Meet Trada, Eunice Choe

Trends in Adolescence Mental Health

- Most common mental disorders in children
  - ADHD 9.8 % ~ 6.0 Million
  - Anxiety 9.4% ~ 5.8 Million
  - Behavioral Problems 8.9% ~ 5.5 Million
  - Depression 4.4% ~ 2.7 Million
- Stats
  - 3 in 4 children who have depression also had anxiety (73.8%)
  - Diagnosed with either anxiety or depression increase
    - 5.4% 2003
    - 8.4% 2012

Trends in Adolescence Mental Health Cont…

- For those that reported a major depressive episode the past year, teens received helped just shy of 50% of the time while 26+ did so 70% of the time



Effects of Untreated Anxiety, Depression, and ADHD

- Detrimental development in young adult hood
  - Impaired mental health
  - Lower life satisfaction
  - Poorer health-related quality of life
- Anxiety can great affect academic performance and attendance
- Suicide
  - $4t^{h}$ leading cause for death for the 15 – 29 age group
- Drug abuse and risky sexual activities

Business Objective

- Want to help facilitate school base mental health prevention
  - Early detection and treatment is critical in improved outcomes
  - Allow for greater access to care
- "Only one-third of schools provide outreach services (including screening to all students)"
  - Help schools determine which students are most likely to need mental health services

SAMHSA

- The Substance Abuse and Mental Health Services Administration 1992
- Agency within the United States
  - HHS Department of Health and Human Services
- Federal Laws related to SAMHSA
  - Affordable Care Act (ACA)
  - Americans with Disabilities Act (ADA)
  - Comprehensive Addiction and Recovery Act (CARA)

Dataset

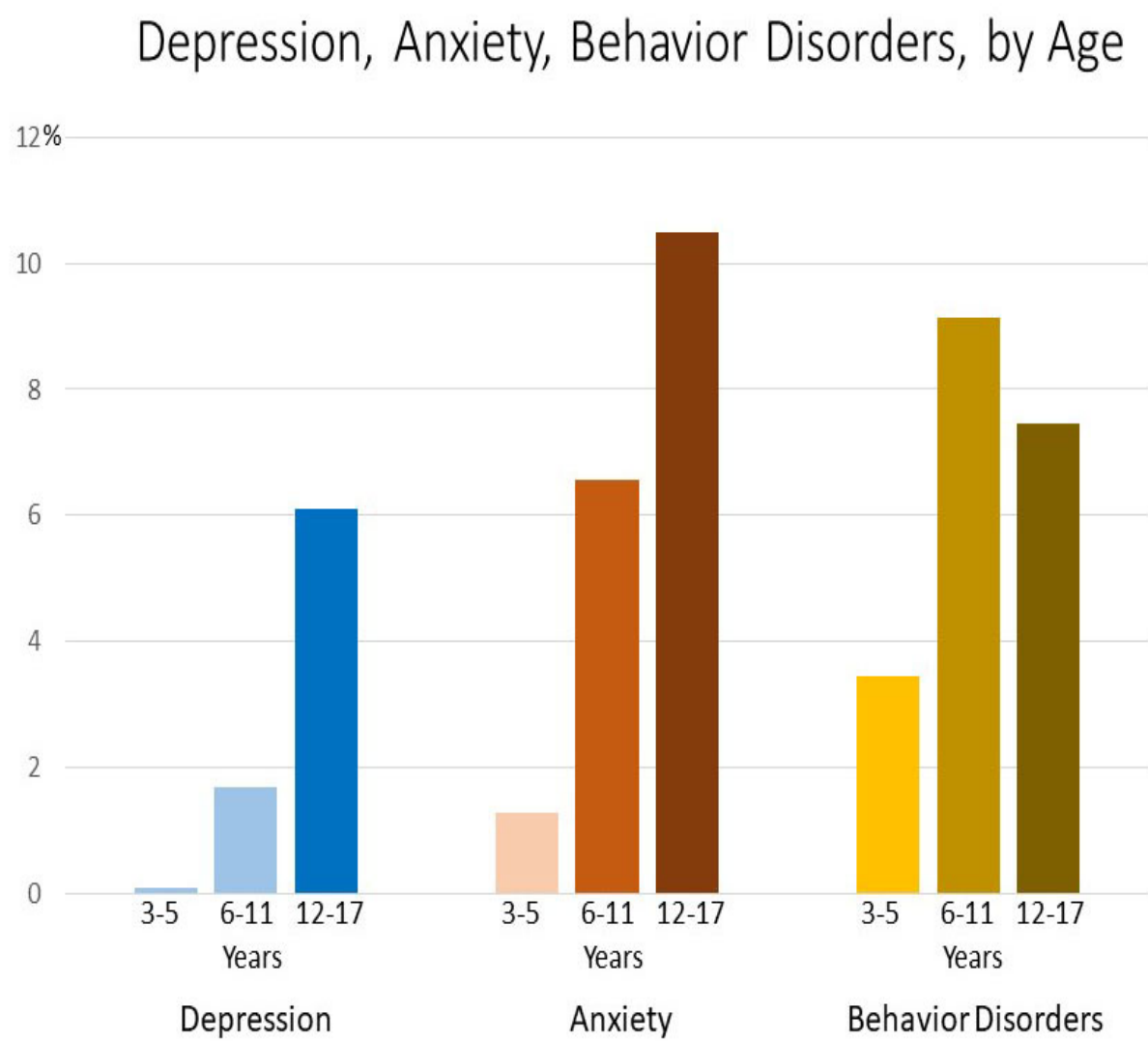
- SAMHSA compiled Client Level Data (MH-CLD) from state mental health authorities relating to demographics and mental health characteristics
  - MH-CLD focuses on individual clients while SAMHSA integrates another dataset focusing on the treatment episodes (MH-TED)
- Original dataset contained 6 million entries with 40 variables
- Does not represent the total demand of mental health services needed
- Disclosure analysis, data swapping, and other techniques were used to protect patient privacy
  - Only small impact on overall trends
  - Most data point stayed the same
  - Special populations (like minorities) were no more affected

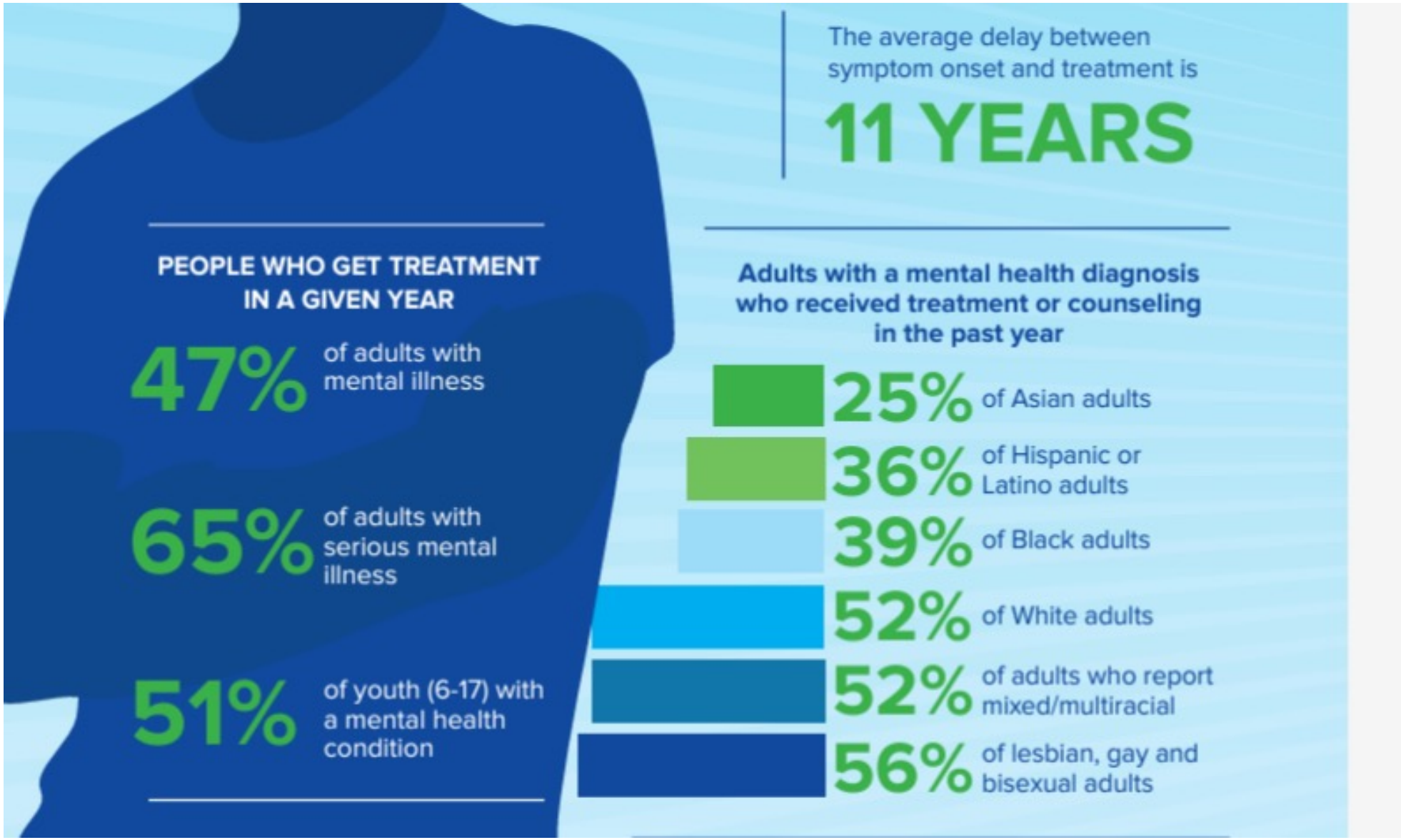
Data Collection/Understanding (CDC)

1. School Health Policies and Practices Studies Questionnaires
2. District level data from schools
3. Exhaustive list of questions collecting data from every dimension about:
    - Availability of Mental Health related services in Schools
    - Policies and measures in place for prohibiting Bullying related acts
    - Resources in place for early intervention
4. Over 400+ variables

Data Mining – Crisp DM

1. **Business Understanding**
   - Understanding MH as a rising challenge
   - Reasons for a rise in MH conditions in minors
   - Understanding the Impact of MH conditions
   - How analytics could prove as a viable solution
2. **Data Collection/Understanding:**
   - Data collection from SAMHSA, CDC
   - 1.8 Million records from SAMHSA (Client level)
   - Supplemental data from CDC on Healthy School Environment/MH and Social Services
3. **Data Preprocessing/Cleaning/Wrangling:**
   - Converting Data to appropriate format
   - Filtering relevant variables based on domain knowledge and statistical tests
   - Cleaning records and imputing for null values
4. **Model Building:**
   - Testing variety of models such as Random Forest Classifiers, XGBoost, Logistic Regression, Decision Trees, Gradient Boosting Machines, LightGBM, CatBoost, AdaBoost to name a few.
   - Trained both single class (3 models one each for unique disorder) and multiclass classification models.
   - As expected, the Single class classification significantly improved precision. More so, as our focus is on detecting each disorder rather than comparing against each disorder.
5. **Testing and evaluation:**
   - 80-20 split for training and testing of models.
   - Stratified splitting to ensure distribution of target variable is the same across both train and test dataset.
   - Cross-Validation for automatic hyper-parameter tuning.
   - Best performing model :
6. **Deployment:**
   - Easy implementation at a Federal level.....


Depression, Anxiety, Behavior Disorders, by Age
12%
10
8
6
4
2
0
3-5
6-11
12-17
Years
Depression
3-5
6-11
12-17
Years
Anxiety
3-5
6-11
12-17
Years
Behavior Disorders


The average delay between symptom onset and treatment is
11 YEARS
PEOPLE WHO GET TREATMENT IN A GIVEN YEAR
47% of adults with mental illness
65% of adults with serious mental illness
51% of youth (6-17) with a mental health condition
Adults with a mental health diagnosis who received treatment or counseling in the past year
25% of Asian adults
36% of Hispanic or Latino adults
39% of Black adults
52% of White adults
52% of adults who report mixed/multiracial
56% of lesbian, gay and bisexual adults

Basic Understanding of Dataset

- Demographics such as Age, Education, Race, Ethnicity, Marital status
- SPHSERVICE- Served in a state psychiatric hospital.
- CMPSERVICE- Served in SMHA-funded/operated community-based program.
- OPISERVICE- Served in 'other psychiatric inpatient center.'
- RTCSERVICE- Served in a residential treatment center.
- IJSSERVICE- Served by an institution under the justice system.
- SMISED- Indicates whether the client has serious mental illness.
- SAP- Substance use problem.
- SUB- Substance use diagnosis.

Data Limitations?

1.Data Scope

MH-CLD offers insights but doesn't capture total national mental health demand. Supplementary data is vital for a comprehensive understanding.

2.Missing Data and Bias

The uneven distribution of missing mental health diagnoses poses a risk of biased prevalence rates in MH-CLD. Addressing this issue is crucial for accurate assessments of the mental health landscape and informed policymaking.

3. Diagnostic Limitations

The allowance of up to three mental health diagnoses per individual in MH-CLD introduces complexities. These diagnoses may not represent a complete enumeration of all diagnoses for individuals served, suggesting potential underrepresentation of mental health conditions.

4.Facility Variations

MH-CLD's data compilation, influenced by state variations in licensing and funding, poses challenges in achieving a standardized and uniform dataset.

Data Cleaning/Preprocessing

- ~6.9 million observations and 40 variables. ~1.8 million records filtered. Further cleaned down to ~1.1 million records.
- Cleaning nulls. Using a rule based + domain knowledge approach. Example : Age groups used to impute for education level
- Dropping variables based on domain knowledge and creating dummies
- Adding Supplemental data for introducing factors that have a significant impact on Mental Health :
    1. Divorced parents
    2. School Violence
    3. Academic Stress
    4. Bullying
    5. Rebelliousness
    6. MH Programs/Facilities
    7. Substance abuse programs
- Randomly chosen records omitted to ensure balanced dataset.
- 4 categories: ADHD Flag, ANXIETY Flag, DEPRESSION Flag, No disorder
- Filtering out 52 additional supplemental variables from CDC questionnaires based on domain knowledge
- Each state provided an average score for the 52 variables based on its Schools' performance. Example: Texas score for number of Counselors on campus : 1.47
- Appending data and converting to flat format

Feature Selection

- Statistical tests of significance:
    - T-test of significance:
        - Null Hypothesis (H0): Assumes that there is no significant difference between the means of the two groups. It posits that any observed difference is due to random chance.
        - Alternative Hypothesis (H1): Contrasts the null by suggesting that there is a significant difference between the means of the two groups.
    2. Chi-square test of significance for categorical variables:
        - Null Hypothesis (H0): There is no significant association between dependent and independent variable.
        - Alternative Hypothesis (H1): There is a signficiant association between dependent and independent variable.
- Final shape :
    - For ADHD – ~690K x 90
    - For Depression - ~640K x 92
    - For Anxiety - ~ 630K x 89

Data Modeling

- Variety of models tested including but not limited to:
  - Random Forest Classifier
  - Gradient Boosting Machines (GBM)
  - XGBoost (Extreme Gradient Boosting)
  - LightGBM
  - CatBoost
  - AdaBoost (Adaptive Boosting)
  - Logistic Regression
  - Decision Trees
  - Naïve Bayes
- Stratified splitting into 80-20% train-test.
- 4-fold cross validation
- Using GridSearchCV to automatically tune hyperparameters
- Random Forest Classifier - Best performing model

Model Evaluation

ADHD

- Best parameters:
  - Max Depth = 20 (Maximum Depth of each tree to 20 levels)
  - Minimum Samples Split = 20 (Each node in each tree will require minimum 20 samples in each node after split)
  - Estimators = 500 (500 individual decision trees)

Results:

	Precision
0	77%
1	72%

Depression

- Best parameters:
  - Max Depth = 20
  - Minimum Samples Split = 20
  - Estimators = 400

Results:

	Precision
0	83%
1	75%

Anxiety

- Best parameters:
  - Max Depth = 30
  - Minimum Samples Split = 20
  - Estimators = 300

Results:

	Precision
0	80%
1	73%

References

- https://www.datafiles.samhsa.gov/dataset/mental-health-client-level-data-2020-mh-cld-2020-ds0001
- https://www.samhsa.gov/about-us/frequently-asked-questions#:~:text=Established%20by%20Congress%20in%201992,behavioral%20health%20of%20the%20nation.
- https://www.cdc.gov/childrensmentalhealth/data.html#ref
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8734087/#:~:text=Mental%20health%20problems%20during%20childhood,quality%20of%20life%20as%20adults.