Project Title: Analyzing Youth-Driven Drug Abuse Patterns in New York City.
Course – CS-GY 6053 Foundation of Data Science, Fall 2023
Instructor: Rumi Chunara, PhD

Authors:
Varad Naik – vvn7114
Sai Nishanth Mettu - sm11326
Praneeth Kumar Thummalapalli - pt2427

Q1) What is the problem?

• Issue: Surging youth drug abuse in NYC, yielding severe social, mental, and financial consequences.
• Drive: Alleviate adverse effects on affected individuals and communities.
• Objective: Develop a holistic strategy encompassing:
1. Predictive Model: Assess youth's susceptibility to drug abuse.
2. Future Forecast Model: Anticipate future cases for effective resource allocation.
3. Clustering Analysis: Identify patterns in social, mental, and peer influences for targeted interventions.

• Method: Draw insights from historical cases, scrutinize social and mental facets, and incorporate financial considerations.
• Approach: Holistic intervention via predictive modelling, future forecasting, and clustering analysis.

Q2) How we learnt the background?

To grasp the background and continue learning, we've employed a multifaceted approach:

1. **Literature Review:**
   - We explored academic journals (e.g., "Journal of Substance Abuse Treatment") for established theories.
   - We reviewed conference proceedings (e.g., "National Conference on Addiction Disorders") for current methodologies.
2. **Analyse Healthcare Reports:**
   - We scrutinized reports from agencies like the CDC and SAMHSA for statistical data and trends.
3. **Conference Involvement:**
   - We attended NYC Gov Health – ‘healthbenifitshome’ conference to stay abreast of the latest predictive modeling advancements in drug abuse.
4. **Continuous Learning:**
   - We stayed informed through ongoing learning initiatives to adapt to evolving research and methodologies.

5. **Citations for the Project :**

   - https://www.nycourts.gov/courts/nyc/drug_treatment/
   - https://www.schools.nyc.gov/school-life/special-situations/substance-abuse-prevention-and-intervention
   - https://nypost.com/2019/12/02/heroin-use-jumps-33-among-nyc-teens-health-department/
   - https://drugabusestatistics.org/teen-drug-use/

   - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5525418/
   - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2681083/
   - https://www.samhsa.gov/data/sites/default/files/reports/rpt39443/2021NSDUHFFRRev010323.pdf
   - https://www.nyc.gov/assets/doh/downloads/pdf/survey/survey-2009drugod.pdf
   - https://www.nytimes.com/2021/11/30/nyregion/supervised-injection-sites-nyc.html
   - https://jamanetwork.com/journals/jamapediatrics/fullarticle/189961
   - https://americanaddictioncenters.org/
   - https://www.foxnews.com/us/nyc-opens-drug-consumption-sites-prevent-overdose-deaths

Q3) What kinds of data have we used?

- **Spatial-Temporal Dataset:**

  - **Spatial Dimensions:**
    - Latitude: Angular distance from the equator (Numerical - Continuous).
    - Longitude: Angular distance from the prime meridian (Numerical - Continuous).
  - **Temporal Dimensions:**
    - Year: Recorded year (Numerical - Discrete).
    - Month: Recorded month (Numerical - Discrete).
  - **Features and Types:**
    - Latitude, Longitude: Numerical (Continuous).
    - Year, Month, Drug_Abuse_Positive, Death: Numerical (Discrete).

- **Individual-level Dataset:**

  - **Temporal Dimension:**
    - No explicit date/time; each row represents an individual.
  - **Spatial Dimension:**
    - Includes "Location" denoting urban, suburban, or rural and specific borough in NYC.
  - **Features and Types:**
    - Numerical (Age, Income, Risk metrics).
    - Categorical (Gender, Race, Employment, etc.).
    - Ordinal (Risk Perception, Propensity, etc.).
    - Textual: Absent.

  - **Scales:**
    - Numerical Scales: Varied scales (e.g., age, income).
    - Ordinal Scales: Implicit in risk-related attributes.

Q4). What Model did we build?

Modelling Choices:

1. **Predictive Model:**
    - Choice: Logistic Regression for simplicity, interpretability, and effectiveness in binary classification.
    - Why Not Others: Avoiding complex models for stakeholder interpretability. Neural networks sacrifice interpretability without significant accuracy gains.
2. **Future Forecast Model:**
    - Choice: ARIMA for time series forecasting due to its ability to capture temporal patterns.
    - Why Not Others: Deep learning models like LSTM networks may be overkill for simpler temporal patterns.
3. **Clustering Analysis:**
    - Choice: K-Means Clustering for identifying patterns in multidimensional data.
    - Why Not Others: Avoiding hierarchical clustering for simplicity. Density-based clustering might be less suitable.

Data Cleaning/Pre-processing:

1. **Systematic Sampling:**
    - Purpose: Create a computationally efficient, representative subset.
    - Benefits: Efficiency, optimization, and preservation of subset representativeness**.**

2. **Data Denoising:**
    - Purpose: Remove outliers and irrelevant columns for imputation and robustness.
    - Benefits: Improved relevance and robustness using Isolation Forest.

3. **Data Imputation:**
    - Purpose: Fill missing values for improved training and unbiased results.
    - Benefits: Improved training, strategy selection, and preservation.

4. **Scaling Columns:**
    - Purpose: Scale 'Income' using Min-Max scaling.
    - Benefits: Equal contribution and normalization.

Model Evaluation Table:

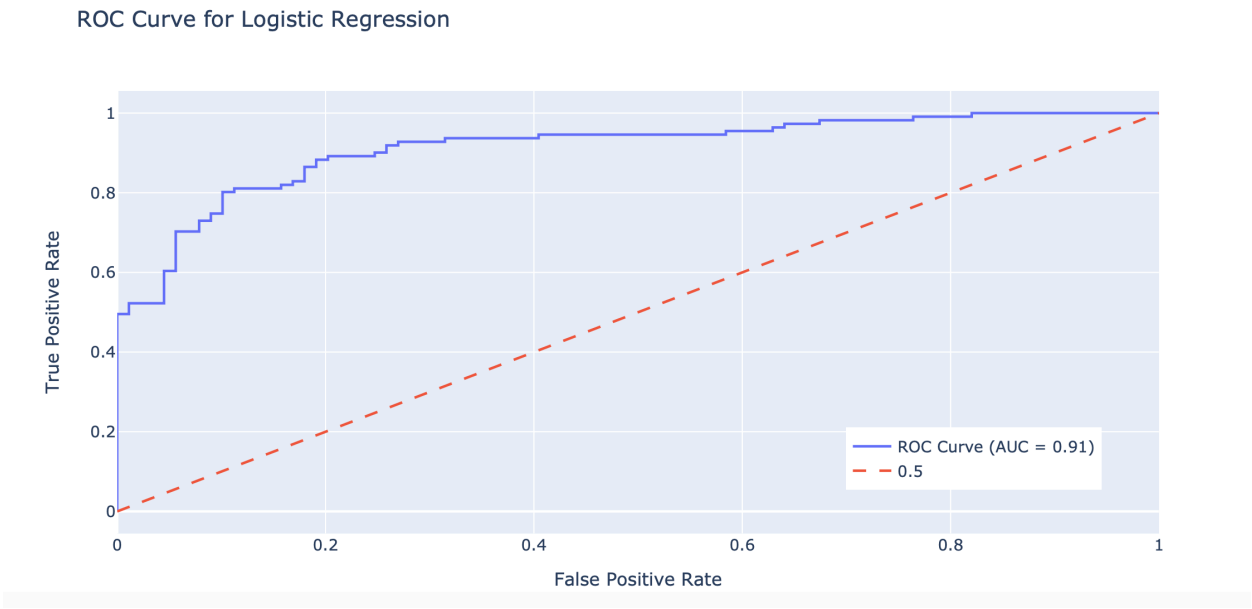
Model	Algorithm/Method	Parameters	Evaluation Metric
Predictive Model	**Logistic Regression**	**Regularization (GridSearchCV)**	**Cross-validated Accuracy, F1, Precision, Recall, Support**
Future Forecast Model	**ARIMA**	**Seasonality Adjustment**	**MAE, RMSE**
Clustering Analysis	**K-Means**	**Number of Clusters**	**Silhouette Score**

This approach balances simplicity, interpretability, and accuracy. Logistic Regression suits binary classification, ARIMA handles temporal forecasting, and K-Means identifies clusters. Data cleaning involves imputation and encoding. Evaluation metrics vary based on model type, ensuring a comprehensive assessment.

Metrics Table:

Model	Type	Result	Evaluation Metric
Predictive Model	**Logistic Regression**	**91% Accuracy 92 % AUC 0.85 Recall**	**Cross-validated Accuracy, F1, Precision, Recall, Support, AUC-ROC Curve**
Future Forecast Model	**ARIMA**	**MSE : 244 RMSE : 15 MAE : 15**	**MAE, RMSE, MSE**
Clustering Analysis	**K-Means & Agglomerative**	**0.60 & 0.58**	**Silhouette Score**

We tailored metrics to each model's goals. For logistic regression, we used accuracy, AUC, and recall measuring correctness, discriminative ability, and specific instance identification. In the ARIMA forecast model, MSE, RMSE, and MAE assessed prediction accuracy and precision. Silhouette Score in clustering analyzed cluster compactness and separation, aligning with our goal of identifying distinct patterns.



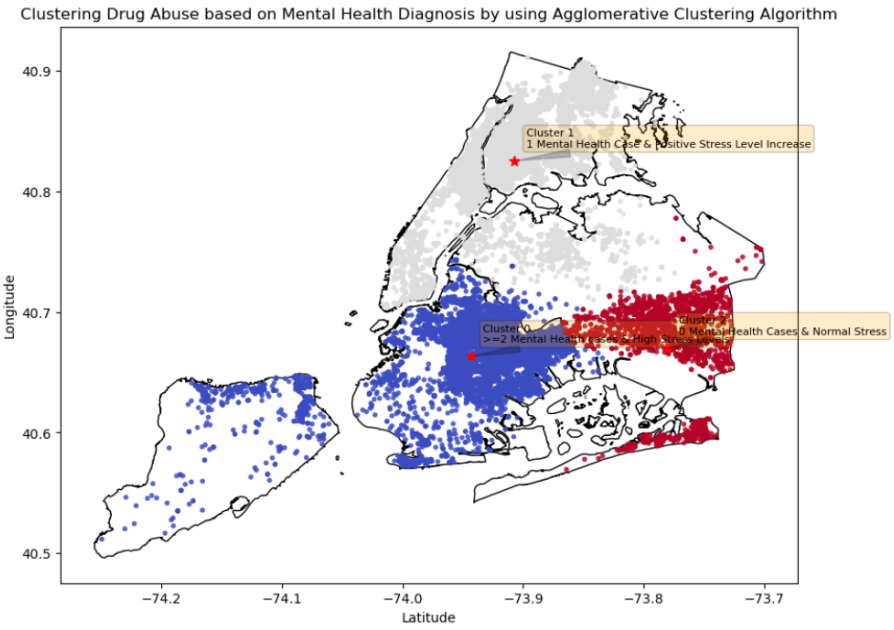


Fig 1. Clustering Drug Abuse Victims based on Mental Health Diagnosis

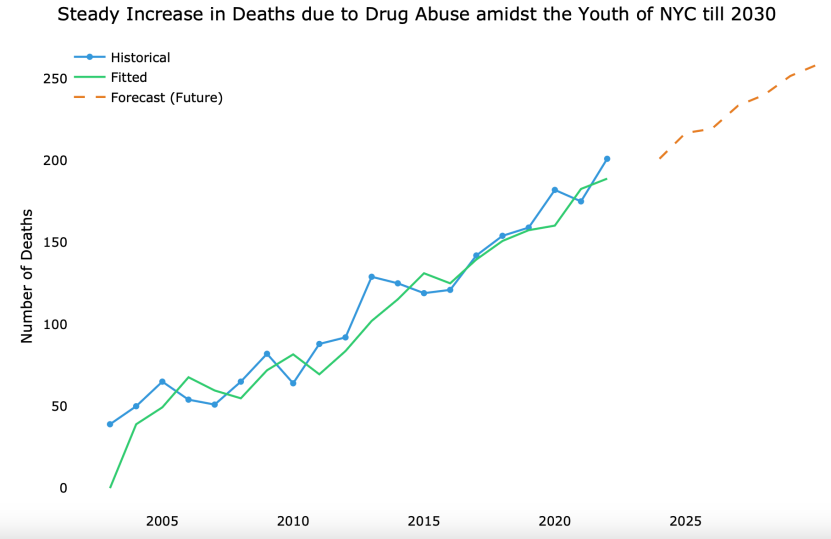


Fig 2. Forecasting the Rate of Growth In Drug Abuse in NYC

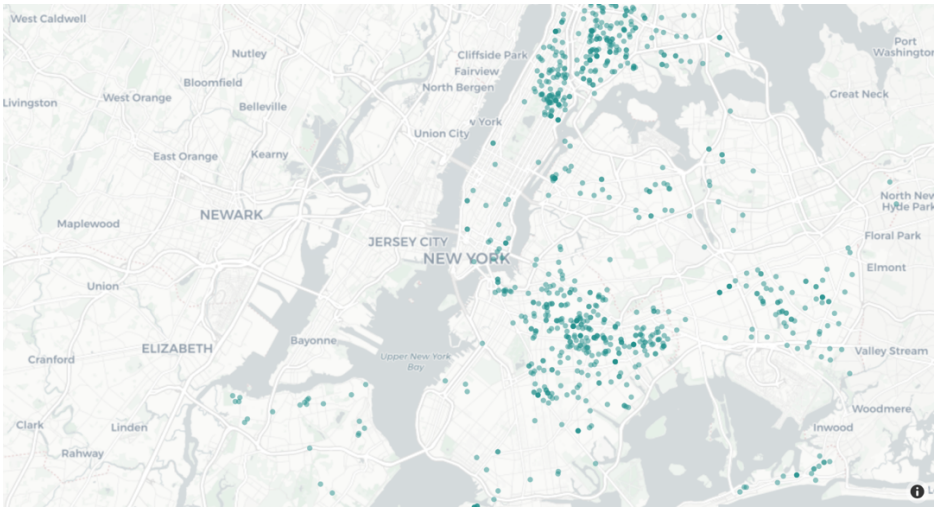


Fig 3. Predicting Drug Abuse Incidence in Youth and visualizing it.

Q5) What assumptions are safe to make?

We confidently assume the reliability of demographic data, grounded in healthcare data standards adherence and well-documented dataset practices. Notably, datasets from reputable sources like NYC DOHMH and CDC WONDER assure completeness, minimizing significant data gaps in drug abuse incidents and mortality data. Assumptions of minimal confidentiality concerns in de-identified data from NYC DOHMH, CDC, and SAMHSA are justified, aligning with robust privacy practices. Additionally, we reasonably assume stable drug abuse patterns during data collection, facilitating meaningful analyses, provided there are no significant shifts in trends that could invalidate conclusions. In public health data, the assumption of independent observations is justifiable, treating incidents as independent events for standard statistical analyses without violating assumptions.

6) What can be the Next steps? :

Enhance the NYC youth substance abuse project by augmenting features with familial and community-level variables. Explore ensemble methods and neural networks for pattern recognition. Refine ARIMA forecasting considering seasonality. Integrate GIS data for spatial insights. Prioritize real-time integration for continuous learning. Assess scalability across cities and demographics. Develop a user-friendly interface with alerts. Explore RNNs and CNNs for advanced analysis. Consider cloud deployment for scalability. Implement API for easy access. Tailor stakeholder interfaces for seamless data flow and real-time predictions.