Enhancing Healthcare's Digital Front Door

PredictNax: Predictive Analysis for Naloxone Distribution

Background

Fentanyl, a potent synthetic opioid, has rapidly escalated into a major public health crisis across the United States, with California being one of the hardest-hit states. Fentanyl's potency—up to 50 to 100 times stronger than morphine—makes it exceptionally dangerous. Even a small dose can lead to an overdose, and its increasing presence in the drug market has exacerbated the opioid epidemic.

San Diego County, a focal point in this crisis, has witnessed a significant rise in fentanyl-related overdoses over the past decade. According to the San Diego County Health and Human Services Agency, opioid overdose deaths have surged dramatically, with fentanyl contributing to a growing proportion of these fatalities. This local trend mirrors national patterns, with fentanyl becoming a leading cause of opioid overdose deaths in the U.S. The rise of fentanyl is partly attributed to its illegal production and distribution, where it is often mixed with other drugs, sometimes without the user's knowledge, increasing the likelihood of an overdose.

One of the critical challenges in this escalating opioid crisis is the growing demand for naloxone (commonly known by its brand name Narcan), a life-saving medication that can reverse the effects of opioid overdoses, including those caused by fentanyl. The rapid surge in fentanyl overdoses has overwhelmed public health systems, placing tremendous pressure on hospitals, emergency services, and community-based naloxone distribution programs to meet the demand for this crucial intervention. Distributing naloxone effectively is further complicated by the geographical and demographic variability in overdose cases, with some areas experiencing disproportionately high numbers of incidents.

In San Diego County alone, public health officials have observed the difficulty in accurately forecasting the demand for naloxone due to the unpredictable nature of fentanyl use and its widespread availability. Traditional distribution models often fall short of addressing this variability, leading to shortages of naloxone in areas where it is needed most.

To address this pressing public health crisis, there is an increasing emphasis on leveraging big data and predictive analytics. By using data science methodologies such as machine learning models—particularly random forest regression models—it is possible to better understand the distribution of overdose cases and anticipate future needs for naloxone. These models can analyze various factors, including historical overdose data, population demographics, socioeconomic indicators, and healthcare access within specific zip codes, to generate more accurate forecasts of naloxone demand. Such data-driven approaches could be instrumental in optimizing naloxone distribution, ensuring that life-saving medication reaches the areas most affected by the opioid epidemic.

By predicting the quantities of naloxone needed in different regions, public health officials can better allocate resources, reduce the impact of overdoses, and ultimately save lives. Addressing the fentanyl crisis requires not only immediate action from healthcare providers and community stakeholders but also long-term, data-informed strategies that allow for a proactive rather than reactive response to the growing epidemic.

Random Forest Model

Methodology & Reasoning

The methodology for this study focuses on creating a predictive model to estimate the number of individuals "at risk" for opioid usage in various communities across San Diego County. This model is designed to inform government agencies on how to distribute naloxone (Narcan) more effectively to those communities most impacted by opioid-related incidents. To build this model, we utilize a random forest regression approach, which allows for robust and interpretable predictions based on a variety of demographic, socioeconomic, and healthcare-related factors.

The independent variables used in this model are sourced from the American Community Survey (ACS) 5-year reports and other publicly available data sources. These variables are chosen based on their relevance to opioid usage risk factors and include:

- **Proportion of adults, seniors, and children:** Understanding the age distribution in a community helps gauge potential opioid usage patterns, as certain age groups may be more vulnerable to opioid dependence.
- **Population 16 and over in the labor force:** Employment status can be a significant determinant of mental health and substance use, as unemployment or unstable work environments often correlate with higher opioid misuse rates.
- **Proportion of females:** Gender can influence the likelihood of opioid usage, with different social and healthcare dynamics affecting males and females.
- Percentage of children in single-parent households and families below the poverty level: These socioeconomic factors are critical, as economic hardship and unstable family environments are commonly linked with increased opioid usage.
- **Employment status:** Employment plays a significant role in mental health and access to healthcare, both of which are relevant in understanding opioid risk.
- Recorded overdose incidents: This historical data on overdose incidents provides a direct indicator of opioid risk in a community and serves as a key feature for the model.
- **Proportion of the population unhoused:** Homelessness has been strongly associated with substance misuse, making this an important variable for predicting opioid-related risks.
- **Number of healthcare facilities:** Access to healthcare is essential for prevention and treatment of opioid misuse, and this variable helps account for the availability of medical interventions.
- **Urban or rural designation:** Geographic context is important, as rural and urban areas often have differing access to healthcare and exposure to opioids.
- **Opioid prescription rates:** High prescription rates are often precursors to misuse, particularly when opioids are prescribed long-term.
- **Drug-related arrests:** This variable helps capture law enforcement data on drug-related activity, providing another dimension of community risk.

The model leverages these variables to train a random forest regression algorithm, which builds multiple decision trees from random subsets of the data. Each tree evaluates different combinations of variables to produce a predicted output, and the model aggregates these results to generate a final prediction for each community. The output variable of the model is the predicted number of individuals at risk for opioid usage in a given community.

By predicting the number of at-risk individuals, the model can be used to estimate how much Narcan should be distributed to each community. This ensures that resources are allocated more efficiently, directing life-saving interventions to areas where they are needed most.

****A random forest model works by building many decision trees from random subsets of the data and combining their predictions. Each tree in the forest makes decisions based on different variables (like opioid prescription rates or poverty levels), splitting the data into smaller groups to predict the outcome—in this case, the number of individuals at risk for opioid usage. The randomness in how trees are built adds diversity, preventing overfitting. Once all trees make their predictions, the model averages them to generate a final, more accurate prediction. This approach provides a robust, data-driven way to estimate Narcan needs across communities.*

Data Collection & Processing

***All data is public and 5-year reports can be accessed through the Census homepage**

- Proportion of adults
- Proportion of seniors
- Proportion of children
- Population 16 and over in labor force
- Proportion of females
- Percentage of children in single parent households
- Proportion of families below the poverty level
- Employment status
- Recorded overdose incidents
- Proportion of population unhoused
- Number of healthcare facilities
- Urban or rural
- Opioid prescription rates
- Drug-related arrests

The data highlighted in yellow was generated using random seed, for the scope of our experiment.

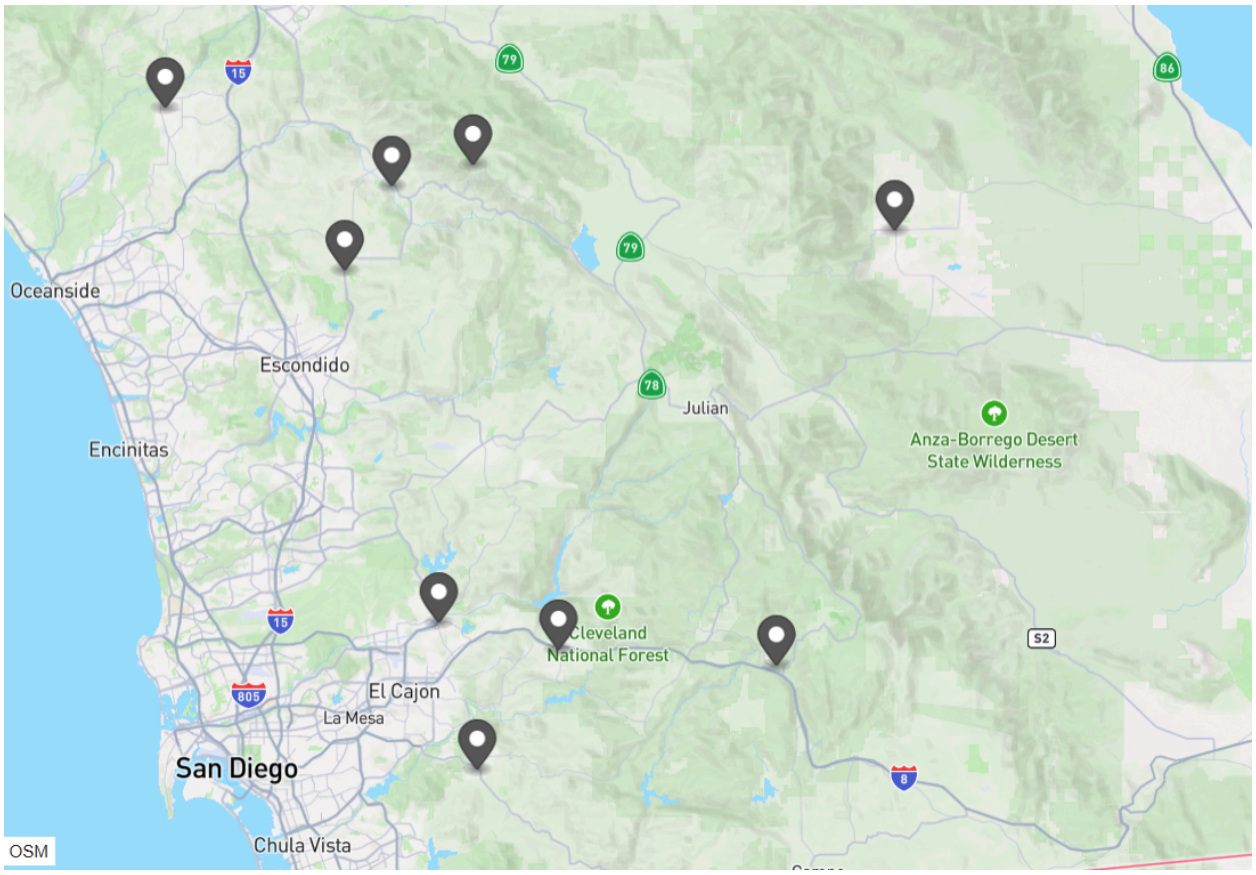
Through using publicly accessible data in the American Community Service 5-year surveys from 2019, then using Python to generate more values. We collected around 2.38 GB of data that, once sorted, contains values on the following variables:

- Proportion of adults
- Proportion of seniors
- Proportion of children
- Labor force participation
- Females in the population
- Single-parent households

- Families below the poverty level
- Overdose incidents

We are treating “community” as it’s own independent variable, and our model will cover the following communities, due to their abundance in sample, “real” data:

- Alpine
- Borrego Springs
- Fallbrook
- Valley Center
- Oceanside-Escondido
- Valley Center
- Jamul
- Pauma Valley
- Palomar-Julian
- Borrego Springs
- Laguna-Pine Valley
- Ramona



*Other variables that were missing, including employment status, number of healthcare facilities, etc., were generated randomly.

Creating Hypothetical Data

Our group chose to generate "fake", yet statistically yet accurate data, as the searching, sorting, and cleaning of accurate, timely data under all these variables in the scope of our project would have led to less time spent on model and methodology creation.

Steps to generate data

1. Statistical analysis of data for key points such as variance, distribution, etc
    - All hypothetical data in our model is only passed through if it fits the general bounds of our 2019 data's minimum and maximum
    - A data set is created for additional years (20xx-20xx), based on the normal distribution of the data – we made sure there are no outliers for replicability
    - The data was loaded into a Pandas dataframe
2. Using Random Seed to reproduce additional data points into our "final" data sheet
   - Our random seed chosen in this experiment is "76"
3. Full code for generating the full, hypothetical dataset with data points on 2019 can be viewed in GitHub here

The "real dataset" for age demographic, sex demographic, and population is a 2019 5-year ACS report – from this real dataset, we created data from 2012-2022 that was generated using random seed. This generated data was then converted into dataframes by year and saved as individual csv files.

The rest of the data points were generated randomly using the following method, with the full notebook available here:

Tomas and I used past research to approximate these rates

```
def generate_random_data(areas, year):
    data = {
        'Community': np.random.choice(areas, size=len(areas), replace=False),
        'Proportion of families below the poverty level': np.random.uniform(0.05, 0.25, size=len(areas)),
        'Employment status': np.random.uniform(0.6, 0.95, size=len(areas)),
        'Recorded overdose incidents': np.random.randint(5, 150, size=len(areas)),
        'Proportion of population unhoused': np.random.uniform(0.01, 0.1, size=len(areas)),
        'Number of healthcare facilities': np.random.randint(1, 20, size=len(areas)),
        'Urban or rural': np.random.choice(['Urban', 'Rural'], size=len(areas)),
        'Opioid prescription rates': np.random.uniform(0.05, 0.3, size=len(areas)),
        'Drug-related arrests': np.random.randint(10, 300, size=len(areas)),
        'Year': [year] * len(areas)
    }
    return pd.DataFrame(data)
```

Lastly, we merged the all generated csv folders made by year

Below is an example of the original data used 2019, next to the random seed generated data set for 2020:

2020 Generated Data

Community	Estimate!!SEX AND AGE!!Total population	Estimate!!SEX AND AGE!!Total population!!Male	Est
San Diego County, California	3458275.7	1680083.44	
Alpine	13558.64	7204.37	
Borrego Springs	2677.42	1306.38	
Camp Pendleton	39286.77	27055.05	
Fallbrook	49289.82	26053.54	
Jamul	20135.19	12084.64	
Laguna-Pine Valley	6021.84	2937.71	
Mountain Empire	7557.44	4347.36	
Oceanside-Escondido	686940.66	348477.87	
Palomar-Julian CCD	5328.42	2814.17	
Pauma Valley CCD	6765.54	3497.81	
Ramona CCD	36149.65	19204.95	
San Diego CCD	2337893.16	1255700.31	
Valley Center CCD	23249.02	12419.45	

2019 ACS Data (original)

GEO_ID	NAME	DP05_0001E	DP05_0001M
Geography	Geographic Area Name	Estimate!!SEX AND AGE!!Total population	Margin of Error!!SEX AND AGE!!Total population
0500000US06073	San Diego County, California	3316073	*****
0600000US0607390030	Alpine	15632	833
0600000US0607390258	Borrego Springs	2706	565
0600000US0607390355	Camp Pendleton	38754	2166
0600000US0607390960	Fallbrook	49232	1489
0600000US0607391440	Jamul	19771	978
0600000US0607391510	Laguna-Pine Valley	5962	516
0600000US0607392030	Mountain Empire	7998	992
0600000US0607392240	Oceanside-Escondido	670537	1963
0600000US0607392350	Palomar-Julian CCD	5268	473
0600000US0607392430	Pauma Valley CCD	6925	785
0600000US0607392550	Ramona CCD	36920	1650
0600000US0607392780	San Diego CCD	2432359	3142
0600000US0607393540	Valley Center CCD	24009	1475

Differences by GEO_ID

2019-2020

San Diego County: 3316073-3458275.7 =

Model Creation

Exploratory Data Analysis (EDA)

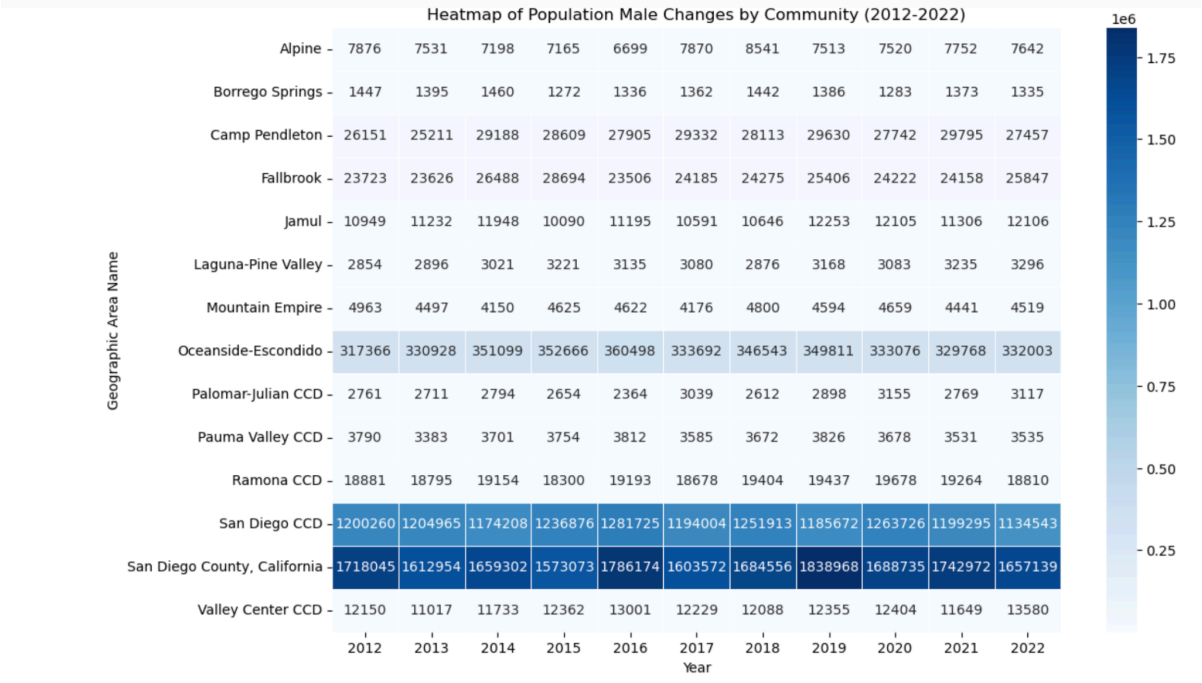
EDA is a crucial part of our process before building the machine learning model, as we are working with demographic and socioeconomic data across multiple years. EDA allows us to understand the distribution of data, detect missing or inconsistent data to ensure that we can address these before modeling, preventing bias and inaccuracies. Additionally, it allows us to identify correlations between variables, and spot trends over time. We can use EDA to remove outliers from our dataset generated by random seed.

Full EDA notebook is available here

In summary:

- We loaded all the datasets and concatenated them into a single dataframe
- We filtered the data and renamed columns
- We checked whether we had any missing values
- Explored the various independent variables

Highlights




Total Population Male Trends (2012-2022)
1e6
Total Population Male
Year
Geographic Area Name
San Diego County, California
Alpine
Borrego Springs
Camp Pendleton
Fallbrook
Jamul
Laguna-Pine Valley
Mountain Empire
Oceanside-Escondido
Palomar-Julian CCD
Pauma Valley CCD
Ramona CCD
San Diego CCD
Valley Center CCD
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2012
2014
2016
2018
2020
2022

Deploying the Model: Random Forest Model

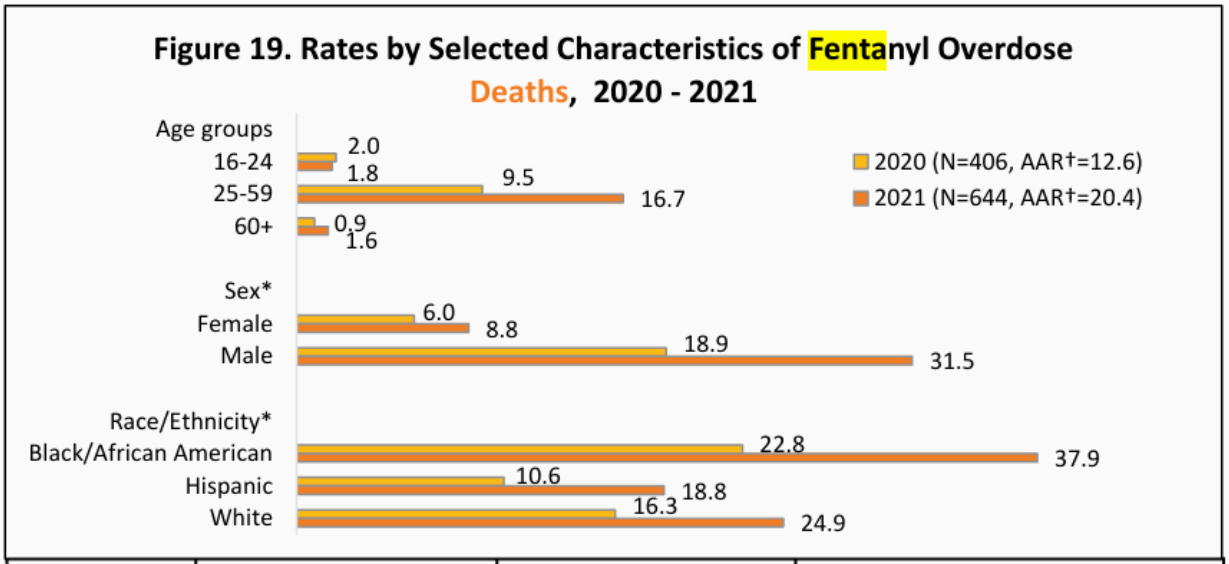
The generated data was loaded into the Python notebook, and we executed the regression available below:

```
Future Predictions for Specific Communities (2023-2025):
                        Community  Year  Predicted Overdose Incidents
0   San Diego County, California  2023                         59.29
1                         Alpine  2023                         72.42
2                 Borrego Springs  2023                         51.47
3                  Camp Pendleton  2023                         67.55
4                       Fallbrook  2023                         65.22
5                           Jamul  2023                         69.42
6              Laguna-Pine Valley  2023                         85.10
7                 Mountain Empire  2023                         73.86
8             Oceanside-Escondido  2023                         84.00
9              Palomar-Julian CCD  2023                         72.97
10               Pauma Valley CCD  2023                         72.95
11                     Ramona CCD  2023                         55.98
12                  San Diego CCD  2023                         82.63
13              Valley Center CCD  2023                         87.24
14  San Diego County, California  2024                         70.23
15                        Alpine  2024                         78.68
16                Borrego Springs  2024                         81.24
17                 Camp Pendleton  2024                         55.58
18                      Fallbrook  2024                         76.79
19                          Jamul  2024                         68.03
20             Laguna-Pine Valley  2024                         71.91
21                Mountain Empire  2024                         72.95
22            Oceanside-Escondido  2024                         69.55
23             Palomar-Julian CCD  2024                         89.74
24               Pauma Valley CCD  2024                         75.35
25                     Ramona CCD  2024                         80.25
26                  San Diego CCD  2024                         69.51
27              Valley Center CCD  2024                        103.30
```

The total predicted amount for 2023

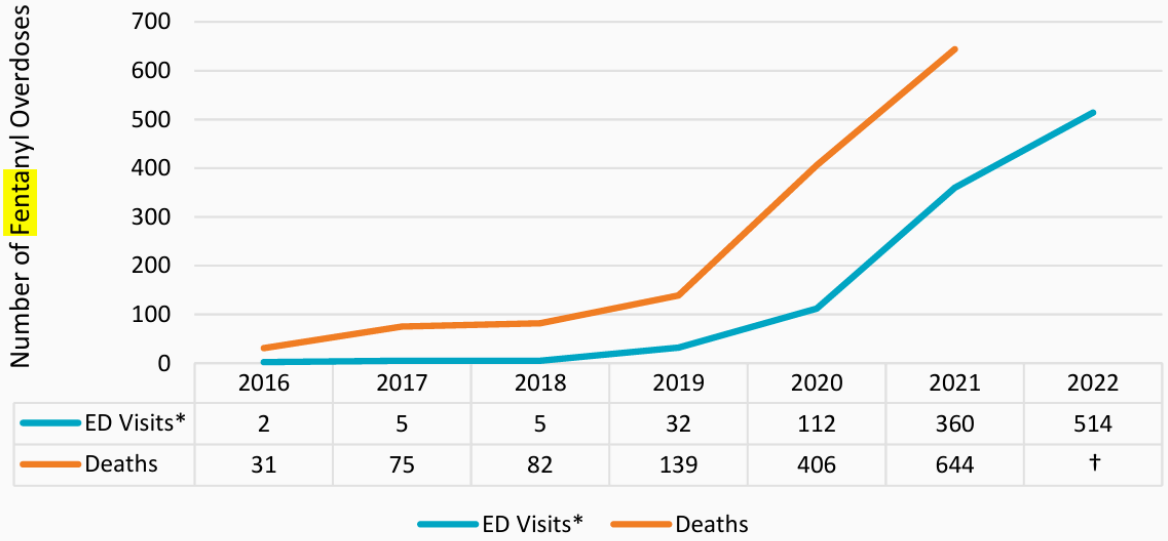
Other Fentanyl Visuals Data:

Fentanyl Overdoses by Demographics



Characteristics	Age groups			Sex		Race/Ethnicity		
	16-24	25-59	60+	Female	Male	Black/AA	Hispanic	White
% Change	-10%	+76%	+78%	+47%	+67%	+66%	+77%	+53%

Figure 13. Fentanyl Overdose Syndromic Surveillance ED Visits* and Deaths by Year



	2016	2017	2018	2019	2020	2021	2022
ED Visits*	2	5	5	32	112	360	514
Deaths	31	75	82	139	406	644	†