Implicit Bias Detection in Policing Practice

Zairan Xiang
University of California, San Diego
zaxiang@ucsd.edu

Mentor: David Danks
University of California, San Diego
ddanks@ucsd.edu

Abstract

Bias in real-world criminal data often leads to serious ethical challenges in the practice of predictive policing using machine learning models. The biased algorithm often reinforces any racial, gender, or other kinds of inequality that it picks up from real-world crime data as it produces biased predictions. The purpose of this research is to explore whether unsupervised machine learning methods can reveal biases in past policing practices in order to avoid feeding the biased data into predicted policing models. To do this, I implemented and used an improved algorithm based on k-means to cluster arrest data to detect any bias pattern implicitly related to protected class attributes like gender and race [1]. The algorithm is successful to detect an implicit correlation between criminals' race and gender with their arrest type, and no biases based on protected personal traits are revealed from the LAPD arrest dataset.

1 Introduction

1.1 Background

Predictive policing help policing practices by applying advanced machine learning models and algorithms and predicting future criminal activity. It has the potential to enable law enforcement to enhance and better direct the patrol resources they have and stop crime before it occurs, thus producing significant crime reduction and providing tremendous economic benefits. For example, predictive hotspot policing enables law enforcement to enhance and better direct the patrol resources they have and stop crime before it occurs according to the predicted geo-spatially arranged maps.

However, there are several major ethical issues with the predictive policing practice. One of the major issues is we are unable to remove all biased data from historical data that we feed into our machine-learning models. We can't do it because human behaviors are complicated, and it is impossible and unrealistic to detect all the discriminatory or biased behaviors in past policing actions. In this case, if our historical data contains biased patterns, our machine learning models will very likely learn this biased pattern and magnify the existing injustice in its predictions.

Moreover, the fairness evaluation of the model might also be impractical with simple methods. Given that crimes are not equally distributed among the population by nature. It is very unlikely that the model will perform equally well across groups given that we would have unbalanced data across groups. Using the false-positive rate to analyze how the model performs with different sub-dataset also doesn't work well. Marginalized groups are often subject to higher false positives, but criminal data hardly capture this. There is no label that indicates if the person is arrested or if the crime is being reported since the dataset only includes arrested or reported cases. Historical criminal data doesn't have a "potential crime" reported case, it also seems unrealistic to have such cases documented in the future. As a result, the false-positive rate might be extremely low: it is very unlikely that the police officers will be able to detect any criminal activity in the region as the model predict there might

[1] https://github.com/zaxiang/Implicit_Bias_Detection

Preprint. Under review.

be, and in reality, criminal activity is really rare. The presence of the police force can also prevent possible criminal activity that we cannot keep track of. In this case, we might not be able to know if our model actually makes the right prediction or if it is being discriminatory over certain groups of people (based on the pattern it learns from the training data) and caused over-policing over those groups.

With the lack of formal fairness criteria in the predictive policing model, unbiasing the historical data is important to solve this ethical issue and avoid replicating or even magnifying any biased human efforts in policing. AI techniques intend to improve the decision-making of law enforcers by adding more efficiency and fairness to the final decisions, so it is crucial for AI to detect the biased pattern in the historical data and be able to make fairer predictions based on fair data.

1.2 Research Overview

In this research project, I introduce an improved algorithm based on k-means with better initializing techniques and a modified loss function with the support of feature reweighting. I removed protected features and used the remaining features to cluster data. By applying this improved clustering method, we should expect a uniform race distribution among clusters that also matched the overall distributions if the data are not overly related to race. Otherwise, we should expect to see the clusters to be highly related to protected personal traits like gender and race if any obvious discrimination based on these traits is present in past policing practices.

2 Literature

Current academic research into AI ethics has faced a lot of challenges. Simply removing sensitive information from the dataset such as gender, age, and race is not enough to remove all biased patterns in the data because data features are often highly correlated with each other: race information might be highly correlated to geographic information based on neighborhoods. One of the most famous approaches to solving the ethical issue regarding biased training data was to generate "fairness metrics" for the models and check if the model can perform equally well across groups (1). However, this method enforces demographic parity, yet the criminal rates vary across groups by nature. The criminal rates are also related to sensitive features such as gender and age by nature.

Another approach that doesn't assume demographic parity calculates the true positive rate and true negative rate for each group. Its measurement of the fairness of the data is based on any inconsistency across groups (2). However, this approach also faces challenges when applied to predictive policing tools and historical crime data. It is often impossible to get true criminal data. For example, for the group with A feature, the true positive rate is Pr{y´=1 | data = A, y=1} where y´ is the prediction and y is the true criminal activity. The true positive rate will calculate the percentage of correct predictions for group A. The y is always unknown. It is certain that y=1 if crimes have been witnessed, yet in most cases, they are not witnessed, resulting in a smaller set of data with y=1. Since the true number of data with y=1 is missing, this fairness measurement is also unrealistic. Moreover, according to Barabas (3), historically marginalized groups are often subject to higher false positives, which makes this approach even more problematic.

As with the debiased model, several works have been done to prevent minority groups in the dataset from being discounted in the machine learning models. For example, Seo et al proposed a cluster-wise reweighting scheme to learn debiased representation and improved on the worst-case accuracy (4). This method might work well to balance the data and perform well with the minority group in the dataset. Yet it can't be applied to predictive policing as crimes are not equally distributed among different groups of people by nature.

3 Dataset

I used the Los Angeles Arrest Data hosted by the city of Los Angeles open data platform [2]. The LAPD arrest dataset reflects arrest incidents from 2010 to 2019 in Los Angeles. It also contains protected information such as the subject's age, race, and gender, which we can use to compare with

[2]https://data.lacity.org/

the cluster results using the unsupervised k-means machine learning model. The dataset contains about 1 million data. Other important features include Arrest Date, Time, and Geographic Areas. Gender and race data are not uniformly distributed: for example, there is about 80 percent of the data is Hispanic and Black Males.

4 Method

4.1 Modified K-Means Model

The k-means is a basic cluster method with a known number of clusters. The traditional clustering is based on the loss function: the sum of squared of the errors (SSE) and assigns cluster labels to each data based on its SSE with the centroid of each cluster. However, the traditional k-means has a very inefficient cluster initialization, which randomly assigns data points as initial centroids for the clusters. A bad initial centroid position could cause the k-means algorithm to converge at a local optimum. In this research, I employed k-means++ which performs much more efficient cluster initialization. K-means++ makes sure that the initial centroids are as far away from each other as possible to avoid the model converging at the local optimum. Specifically, k-means++ selects each new centroid based on the maximum squared distance from the rest of the centroids.

The modified k-means algorithm also supports feature reweighting by adding penalties to the loss function. The data features contain categorical variables such as arrest type and charge type. Performing one-hot-encoding on the categorical variables would increase its number of dimensions thus inevitably adding more weight to the corresponding features, thus resulting in unbalanced weights across features. In order to avoid the model from paying more attention to categorical variables such as arrest type than numerical variables such as latitude, I improved the loss function by modifying the distance function:

- The original SSE loss function using Euclidean distance

$$d\left(p,q\right)=\sqrt{\sum_{i=1}^{n}\left(q_i-p_i\right)^2}$$

- the modified SSE loss function with the added penalty

$$d^{'}\left(p,q\right)=\sqrt{\sum_{z=1}^{Z}(\frac{1}{\alpha_z}\sum_{i=1}^{n}\left(q_i^z-p_i^z\right)^2)}$$

  where Z is the number of feature sets we have, and α_z is the dimensions of the feature for each z. For example, if data $\{q_i\}_{n=1}^{N}$ has 4 dimensions, with the first 3 belonging to a single feature and the 4th one belonging to a feature, Z would be 2 here, and α_1 would be 3.

The added penalty α would ensure the model pays equal amounts of attention to each feature and not overlooking numerical features that only have 1 dimension. The next step I did is normalize all dimensions to range from 0 to 1 to further make sure that the features with large numbers like location (latitude and longitude) would not play a more dominant role than others.

With the support of features reweighting and a more efficient cluster initialization, the basic idea of the modified k-means is as follows in Algorithm 1: First iteratively select centroids for each cluster using k-means++ to make sure the centroids are far away from each other. For each data point, calculate its distance with each centroid and pick the one with the shortest distance (minimal SSE) as its cluster assignment using the modified Euclidean distance with the added penalty. Next, recalculate and update the centroids for each cluster according to the new cluster assignments in this iteration. The process will be done recursively until cluster assignments for all data points stay the same as the previous iteration or it reached the maximum iterations.

The k-means-based algorithm would put data into different clusters based on their vector similarity. To determine the number of clusters k, I used the elbow method to determine the optimal number of clusters and plot the number of clusters k vs. model fit (SSE). I then picked the elbow of the curve as the number of clusters to use.

After picking the optimal number of clusters, I then performed Cluster analysis with the optimal k. The first step is to remove any protected features gender and race from the data. By applying the modified k-means algorithm with optimal k, if other data features are not highly related to race and

Algorithm 1 K-means Algorithm

Input Data vector $\{x_n\}_{n=1}^{N}$, number of cluster K

$\{\mu_k\}_{k=1}^{K} \leftarrow$ **k-means++** $(\{x_n\}_{n=1}^{N}, k)$ ▷ Use K-Means++ to initialize centroids
while max iteration **do**
  for $n \leftarrow 1...N$ **do** ▷ new cluster assignment based on distance for each observation
    $r_n^{'} \leftarrow min_k$ Modified_SSE $(x_n - \{\mu_k\}_{k=1}^{K})$
  end for
  for $k \leftarrow 1...K$ **do** ▷ update centroids based on current assignment
    $\mu_k \leftarrow mean\{x_n | r_n^{'} = k\}$
  end for
  if $\{r_n^{'}\}_{n=1}^{N} = \{r_n\}_{n=1}^{N}$ **then**
    Break ▷ early stopping
  end if
end while

Output cluster assignments $\{r_n\}_{n=1}^{N}$ for each data point, and k cluster centroids point $\{\mu_k\}_{k=1}^{K}$

gender, each cluster should have approximately the same race and gender distributions, and they should also stay the same with the gender and race distributions in all arrest data. This is because the k-means algorithm groups data based on their similarity. data would be similar to each other if they share information on gender or race, and we would expect similar data to fall into the same cluster.

5 Result

I run the model on arrest data with normalized features day, month, hour, minute, area of arrest, latitude, longitude, arrest type, charge type, and age, from 3 clusters to 6 clusters and plotted the SSE with K. From Figure 1, the elbow point is at 5 clusters, thus I set K to 5 for the next step which is clusters analysis.

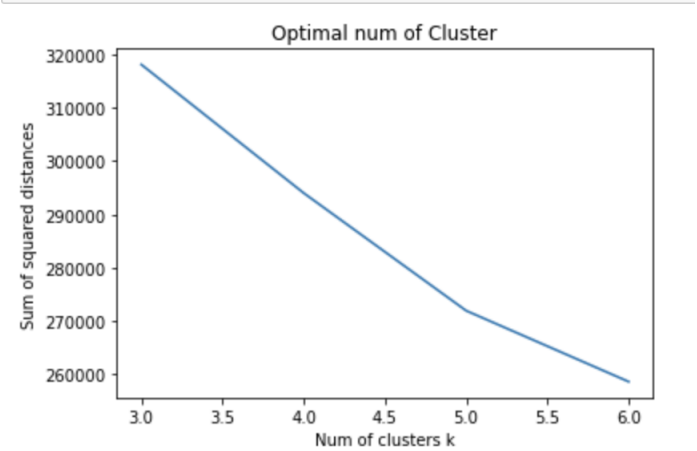


Figure 1: Optimal Number of Clusters K

Next step I applied the modified k-means model with k=5 on arrest data and plotted the cluster assignments with dimensional reductions in Figure 2. From the graph, we can see some pretty clear clustering. The clusters are not perfectly grouped which was expected because by applying dimensional reduction, all clusters' information has been compressed into 2 dimensions.

In Figure 3, the red bar represents the percentage of certain races across all data, and the dark and light green bars represent the percentage of certain races within each cluster. Data contains races group: {A Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z -

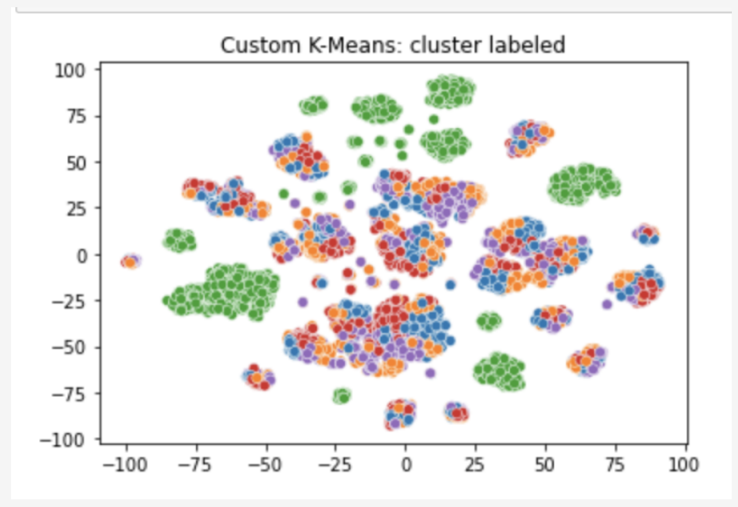


Figure 2: Cluster Assignments

Asian Indian}. The plot only shows four race groups that have a percentage over 1% in the dataset, which are Black (B), Hispanic/Latin/Mexican(H), White (W), and Other (O).

By plotting the race distribution within 5 clusters and comparing it with the overall race distribution across all clusters, we can see from Figure 3 that for the middle cluster, the race distribution shows very different from the other clusters. The total variation distance (tvd) between racial distribution in cluster 2 with the overall racial distribution is 0.2, while the tvd for the other four clusters stays below 0.1. I then looked further into this cluster to check on possible reasons why it has such disproportional distribution.

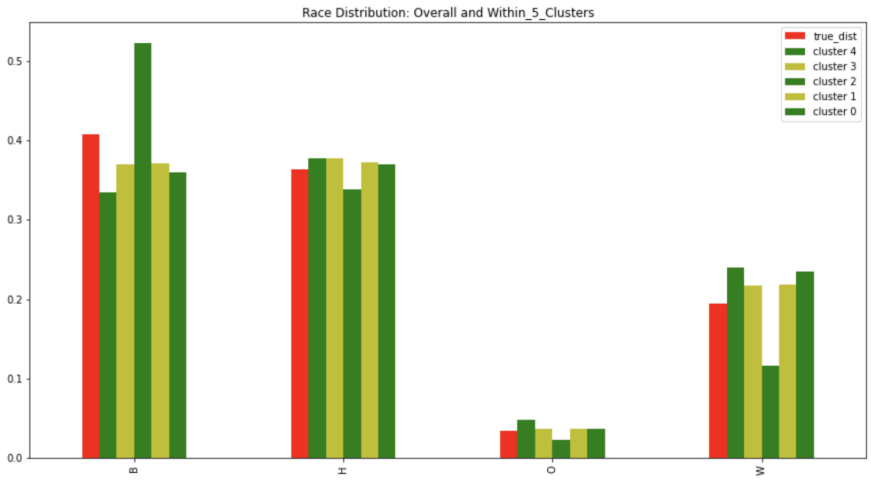
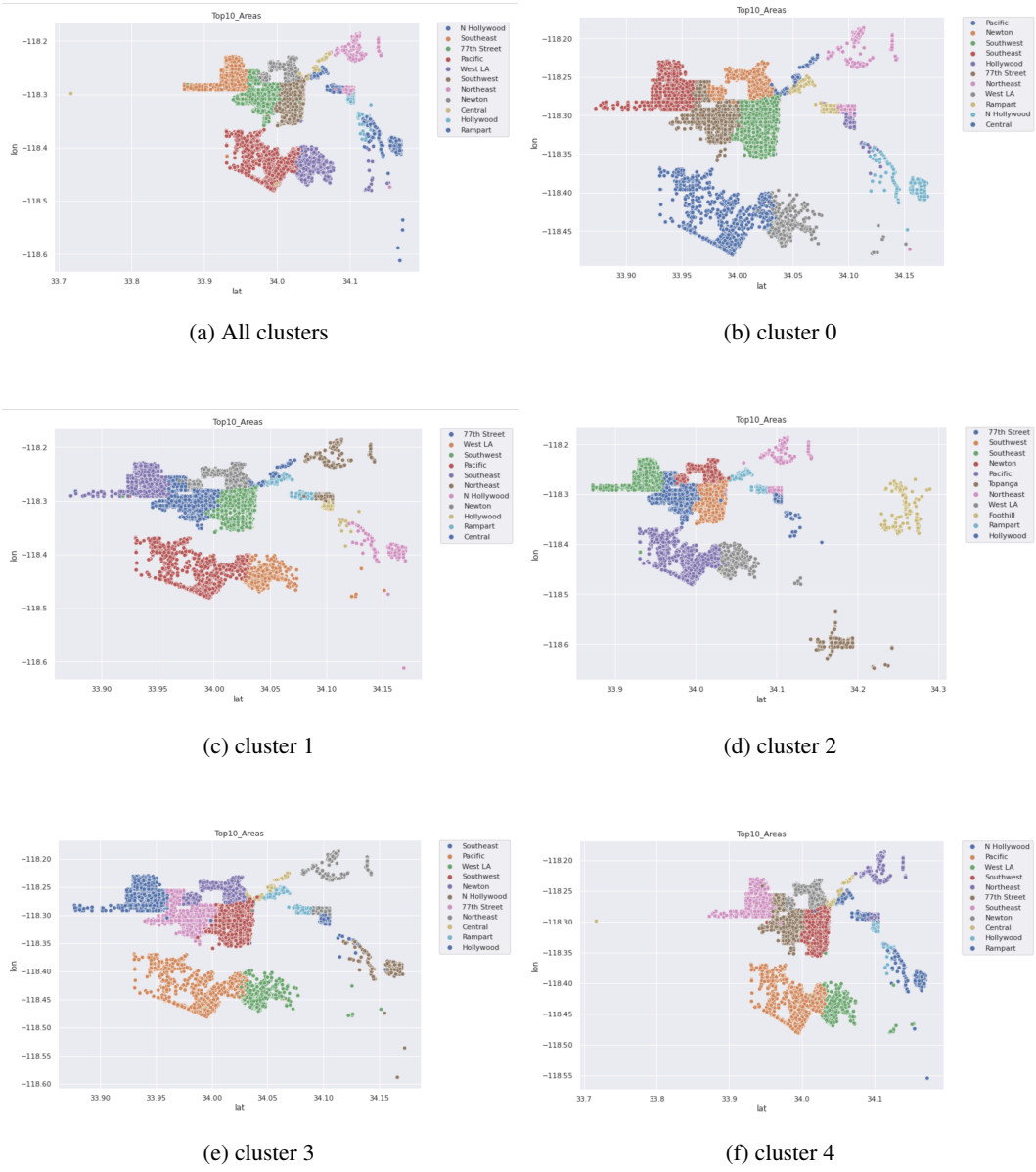


Figure 3: Race Distribution Among Clusters

With the arrest incident in cluster 2, the data contains more Black arrest cases and fewer White arrest cases compared to the true racial distribution in the dataset. As shown in the table in Figure 5, the arrest incidents contain the most data from 77th Street data, while the area with the largest incident cases in the other four clusters is to be Pacific instead. Moreover, as shown in Figure 4d, cluster 2 is the only cluster that contains data from Foothill and Topanga (the far right data points on the map), which both have large proportions of White residents. The inclusion of the neighborhoods from the Foothill and Topanga areas might tell some interesting information since Foothill and Topanga shows opposite racial distribution compared to what it has in cluster 2. We can also see that except

for cluster 2, the other four clusters share similar geographic information (with cluster 4 having an outlier), which is expected and matched the racial distribution information in Figure 3.



(a) All clusters

(b) cluster 0

(c) cluster 1

(d) cluster 2

(e) cluster 3

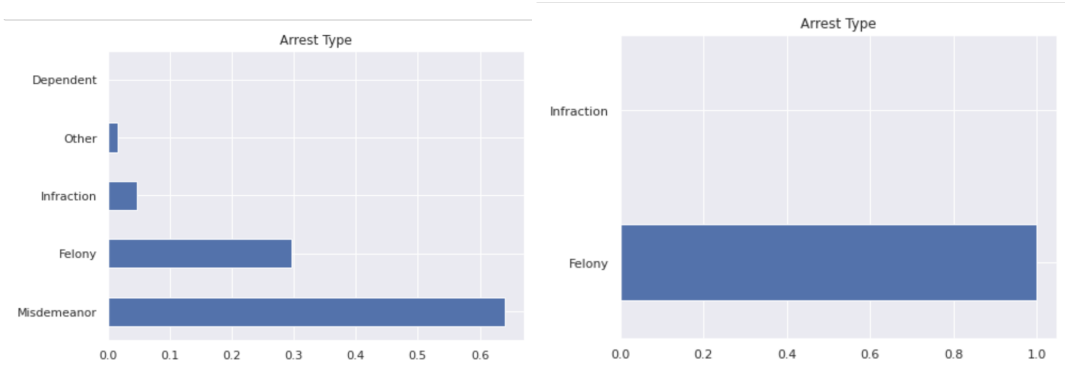
(f) cluster 4

Figure 4: Geographic Information of Arrest Incidents in Each Cluster

By looking into the centroid for cluster 2 and comparing it with the other four centroids, it seems obvious that the arrest type is the dominant feature in the clustering: with the centroid in cluster 2 highly leaning toward the Felony arrest type. As shown in Figure 6 with a comparison with overall arrest type distribution, cluster 2 only contains almost only Felony incidents. We can conclude that racial information is highly related to arrest type in LAPD criminal data. The geographic map might also reveal some suspicious information based on the opposite racial distribution in criminal data compared to the demographic information in the residence.

	all_data	most_area_0	most_area_1	most_area_2	most_area_3	most_area_4
0	Pacific	Pacific	Pacific	77th Street	Pacific	Pacific
1	Southwest	Southwest	Southwest	Southwest	Southwest	Southwest
2	77th Street	77th Street	77th Street	Southeast	77th Street	77th Street
3	Southeast	Southeast	Southeast	Newton	Southeast	Southeast
4	Newton	Newton	Newton	Pacific	Newton	Newton
5	West LA	West LA	West LA	Northeast	West LA	West LA
6	Northeast	Northeast	Northeast	Rampart	Northeast	N Hollywood
7	Hollywood	Hollywood	Hollywood	West LA	Hollywood	Hollywood
8	Rampart	Central	Central	Hollywood	Central	Northeast
9	Central	Rampart	Rampart	Foothill	Rampart	Rampart
10	N Hollywood	N Hollywood	N Hollywood	Topanga	N Hollywood	Central

Figure 5: Most Incidents Areas in each Cluster



(a) Arrest Type in All clusters

(b) Arrest Type in Cluster 2

Figure 6: Arrest Type Distribution

6 Discussion

6.1 Gender Feature

I discussed the cluster analysis on the racial aspect from the results of k-means clustering in previous sections. The gender distribution among clusters doesn't show large inconsistencies with the overall gender distribution. however, the inclusion of gender information in the model shows that gender is also highly related to arrest type.

Using data features day, month, hour, minute, area of arrest, latitude, longitude, arrest type, charge type, age, gender, the racial distribution within all 5 clusters shows a significant difference with the overall true racial distribution in all data, as shown in Figure 7. The gender feature plays a dominant role in the clustering process as each cluster only contains one gender: the data were perfectly grouped according to gender information. The arrest type distributions in 5 clusters also follow the same patterns as we've already seen in the previous section: some clusters are dominated by Felony incidents data. With the inclusion of the Gender variable, I also spotted that the arrest type is also highly correlated to gender, with almost all Felony data falling into the Male category.

The cluster assignments graph with gender variable included also confirms the correlation between Gender and Arrest Type. As shown in Figure 8 with a comparison from cluster assignments without gender information in the model on the right, we can see that the clustering with gender included in

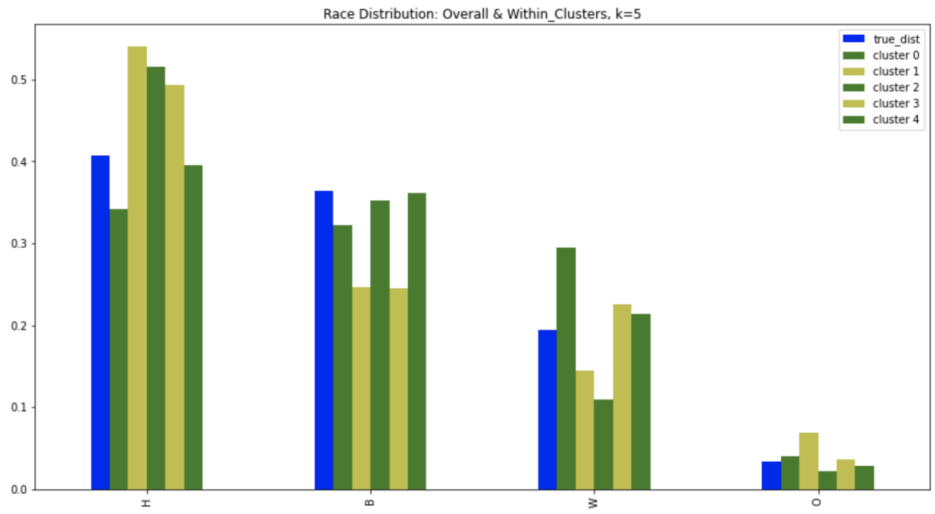


Figure 7: Racial Distribution Within Cluster (with Gender Feature)

the model turns out to perform better, with clearer groups. The better performance in clustering data on the left is caused by the added duplicate information that gender and arrest type share. However, in this research, we would want to remove gender information since we don't want to include protected features or any duplicate features in our model. Moreover, the modified k-means algorithm always stops at early iteration when the gender feature is included, but the algorithm would reach maximum iteration every time the gender feature is not included. This is also expected, the addition of duplicate information in the feature would help the model in clustering the data significantly, resulting in much faster and better performance.

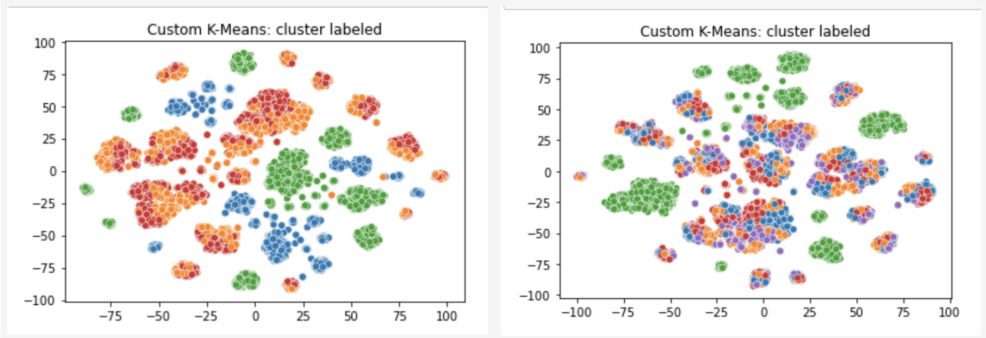


Figure 8: Data Cluster Label (w/ Gender on Left, and without Gender on right)

6.2 Compared Methods

I also compared my improved algorithm of k-means with the k-means model from sk-learn. the k-means package imported from sk-learn doesn't give much flexibility to play with it, and I used it as a baseline model before implementing my improved k-means algorithm. The k-means model from sk-learn performed poorly compared to my modified algorithm as shown in the data cluster graph in Figure 9 with no clear data groupings.

6.3 Limitation and Future Work

Model's potential: In this research, the algorithm aims to pay less attention to one hot encoding categorical feature to ensure equal weight among features, but the modified k-means algorithm can also be helpful for future uses that train the model to pay special attention to the biased feature.

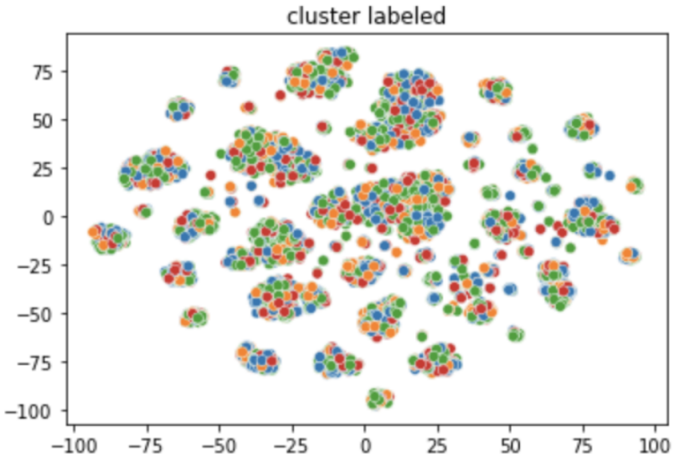


Figure 9: Data Cluster Label (Sklearn K-means model)

Limitation: There is still a large limitation in using clustering techniques to detect implicit bias in the data as it can only capture some significant observations on the suspicious correlations between data features. We cannot move on to more complicated and further analysis using clustering techniques other than saying some correlations are suspicious at this time. An example of the suspicious correlation would be the opposite racial distribution between cluster 2 with the demographics information in the areas that cluster 2 covers. Hypothesis could be made based on this that unfair policing practices might be present in the neighborhood that have a large proportion of White residents. However, the clustering techniques used in this research are not enough to test this hypothesis.

Conclusion: According to the results from the cluster analysis, we found some interesting correlations between criminal activities with criminals' gender and race, yet there is not enough evidence showing the presence of biased policing practices based on criminals' protected features. Again, it is still not certain whether or not the dataset contains other kinds of biases since the true distributions of crime are unknown. Some future work might be to employ more advanced methods to test out some possible hypotheses we conclude from this research.

Future work: future work after the bias detection is to find a "better" model that can debias the data. The evaluation of the predictive policing model is also challenging. since we don't actually have the true criminal activity, it's gonna be a challenge to evaluate or test the model. Some possible methods might involve splitting the data into days or hours and using the data on day 0 to day 1 as training data (X), and data on day 2 as the results (y), the modal prediction would be y'. By doing this, we could evaluate the model for each protected group of race, sex, etc.

7 Github Page

https://github.com/zaxiang/Implicit_Bias_Detection

References

[1] Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

[2] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).

[3] Barabas, Chelsea. "Beyond bias:“Ethical AI” in criminal law." (2020).

[4] Seo, Seonguk, Joon-Young Lee, and Bohyung Han. "Unsupervised Learning of Debiased Representations with Pseudo-Attributes." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.