**Title page**

*Hello, I’m Dylan Rohan and today I’ll be presenting my research project*

*‘Clustering Algorithms as a Means of Identifying   
Patterns of Bias’*

**Contents**

I will be covering the Problem description, Overview, Methodology, Results and finishing up with conclusions and a brief demonstration.

**Problem description**

Bias is almost inescapable in any domain, and it can lead to inefficiency, mistrust, and injustice. Over the past few years these issues have become more obvious as machine learning tools learn these behaviours. (Miller, Peek & Parker 2020)(van Giffen, Herhausen & Fahse 2022), which was well represented in the Netflix documentary ‘ the social dilemma’ but has also been seen on dating apps, social media, HR tools, loan approvals (Bivens & Hoque 2018; Garcia, Garcia & Rigobon 2024; Köchling & Wehner 2020; McDavid 2020; Nader 2020; Narr 2021).

* **Incentives**

While there are frame works that can measure biases organizationally, such as the Implicit bias association test, go/no-go association test, implicit relational assessment procedure, HR protocols and the like, they tend to be ‘perpetrator’ focussed and have many negative factors associated with them.

But addressing bias has many benefits beyond the obvious ethical reasons.

* **Research questions**

In this project, I posit the idea that if machine learning tools have shown an ability to detect the patterns of bias, then perhaps machine learning can offer a better alternative than current methods. Specifically, my research questions are:

Can clustering algorithms be used to identify potential patterns of bias?

And if so, can it be done in a way that is accessible both conceptually and computationally?

* + Selected reference domain

The concept was conceived when looking at this dataset from the American medical system which was trying to determine the causes of readmission. It provides identity features, as well as treatment or experience features. It has over 100,000 instances with 48 features and all the appropriate licensing. It can be found on the UC Irvine machine learning repository with this link.

* + Selected features

Pause here for more information

**Overview**

Essentially, this project aims to use a **pre-existing dataset that potentially capture biased patterns indirectly and** is unlikely to have been tampered with or to ever be expected for use for this purpose offering some protection from obfuscation. It attempts to find groups using **Clustering algorithms** withthe latent information and feature interactions hidden in the data.

The aim is **not** to prove a bias. It is to:

Identify groups based on patterns from historical data

Assess those groups to determine if there is an association with an identity feature.

If there is an association, promote productive conversations that are ‘victim’ focussed to determine the reason.

**Methodology**

Quite a lot of trial and error went into this, but I’ll briefly step through each phase

* Phase 1

Firstly some simple data cleaning tasks revealed many missing values and blank features for columns I would have liked such as weight to investigate fatphobic potential patterns, and patient identifying numbers. Without the later, I had to interpret each row as a hospital experience, and not an individual patient. Domain and context research also consumed many hours.

* Phase 2

An EDA, with univariate and multivariate analysis was conducted, along with outlier observations. From here, features were selected, and the datasets were pre-processed.

As briefly seen before there were four datasets constructed from largely the same features. They included a high instance, low instance, PCA transformed and UMAP embedded. Initially this was to include two features in the low instance set, but as the project evolved the low instance set was abandoned as it didn’t mimic a real world setting very well.

* Phase 3

Here dimensionality techniques were used to capture patterns and reduce processing time with viewer dimensions. Once all this was completed, several clustering algorithms were assessed.

* Phase 4

In the final phase, each model was assessed and some tools and meaning were extracted.

**Results**

* Exploratory Data analysis

The EDA didn’t reveal anything too surprising. Just things like older people spend more time in hospital and that there are many outliers in each feature. I decided to include outliers as some of the story may be hidden there, but am unlikely to include anything about this phase in the final report.

* Principal component analysis

The PCA works by looking at orthogonal dimensions and sequentially captures variation. It’s like taking photos of the solar system that maximises the seperation between planets before heading to another viewpoint that takes a picture of the second most separation, and so on. The pictures here represent the components. It was found that 14 components could capture 95% of the variance, however as you can see in the plots. This transformation doesn’t make it obvious that clusters will be found. Most features here are sitting close the origin, and only a few features seem to be important to each component.

* Uniform Manifold Approximation and Projection

The UMAP embedding was far more clumpy. This model is mathematically heavy, but essentially converts the dataspace into a topological space, throws a blanket over the top, and then uses some cheeky mathematics to approximate the local structures and global structures. This embedding was able to take 45 features, embed them into two and then return to 45 with RMSE of 0.22. That’s effectively a 20% loss on a two-way trip from 45 to 2 and back, however the local structures have been preserved and UMAP is not supposed to be able to reverse transform exactly. It is simply the only metric available to meaningfully validate the embedding. This method is very effective on large datasets and is more efficient than other reduction techniques such as T-SNE, its also proved itself when used in supervised settings.

* K-means

k-means is a distance-based algorithms that uses proximity to points in the dataspace to try and group data, but it’s not very effective with noise or non-spherical shapes. Silhouette scores were used to assess the findings across five folds for each attempt. The UMAP embedding was the only one to have any potential utility, but expecting users to know a reasonable K value is unrealistic and this model would take far too much time to run. Additionally, its unlikely that spherical structures would exist.

* DBSCAN

DBSCAN operates by finding pockets of density, and has the advantage of being insensitive to noise. However, the tuning phase is time consuming, and assumes even density in clusters. This model was unable to present an output as memory issues results in the model crashing the kernel 44 hours in.

* HDBSCAN

HDBSCAN is a very elegant model that creates something akin to a probability density function and uses the topology of the space to determine dense pockets. It is insensitive to noise, requires fewer parameters to be tuned and is far more computationally efficient. Due to time constraints it was only tested on the UMAP embedding. When I visualized each cluster, I was satisfied with the boundaries being made, it only labelled around 15% of the data as noise, and the silhouette scores from each fold had minimal variance.

* Statistical Analysis

In the analysis phase, the identity features were assigned to their given cluster and chi-squared test of independence, crahmare’s V and z-tests were then used to determine that there was a significant low to moderate association between gender, age, and ethnicity with their label, and that many of these clusters were more than 3 standard deviations from the expected value to a (p-values < 0.05). From here, Domain experts would need to assess these groups and determine whether the experience associated with that cluster is justifiable or what could be done to improve this experience.

**Conclusion**

Something I wish I had considered more prior to starting this dataset is the fact that treatments have changed over the last decade. One treatment that was once the gold standard may now only be offered to a lower socioeconomic group or a victim of bias. This might result in some clusters having old privilege and new-underprivileged experiences appearing together which obfuscates the findings here. If I were to repeat the project, I would use a smaller time frame for one specific region.

In ters of progressing the field, I could only find one paper which clustered language to determine how a bias may influence natural language processing applications (Caliskan et al. 2022). Other discussions are predominantly regarding clustering for segmentation analyses or similar, and how this can result in biased outcomes (Klein 2016; Nakip, Gökmen & Mohammed 2017). Fewer still comment on the patterns of bias being identified inadvertently, be there as it may (Recchia et al. 2022). There also does not appear to be any literature presenting clustering as an application to directly identify potential bias as has been attempted here.

The UMAP+HDBSCAN combination has the capacity to capture patterns of bias in a more efficient way, both organizationally and computationally. Itis vaguely explainable and results in healthier conversations than other methods

**Demonstration**

I will quickly run through the code, and demo one of the tools I’ve put together.

I’ve annotated everything so feel free to pause as we’re going through to read my explanations.

* **References**