Clustering Algorithms for Identifying Patterns of Bias: A Novel Approach

Abstract:

Tools used in the identification and measurement of bias are limited, costly, and may result in conflict. In this paper, clustering models attempted to identify patterns of bias hidden in latent information in a manner that is low cost, fast, and of low conceptual complexity. PCA and UMAP dimensionality techniques were used in conjunction with K-means, DBSCAN, and HDBSCAN. The combination of UMAP and HDBSCAN met these goals, providing an output that promoted collaboration, further investigation, and was non-accusatory as it focuses on patterns and not perpetrators. This combination identified clusters at a very granular level. The identified clusters were statistically associated with identity features, with several deviating beyond the expected value. This demonstrated the utility of the method and implied that clustering algorithms can be used in this manner to assist in the identification of bias.

Keywords: Implicit bias, systemic bias, Clustering, HDBSCAN, UMAP, behavioural science

Introduction

The advancement of machine learning and artificial intelligence has introduced thousands of tools that competently reach their primary objective, with their creators handling unexpected issues as they occur. These tools, built in an imperfect world, are sometimes unintentionally built on bias or false assumptions, or learn a bias over time (van Giffen, Herhausen & Fahse 2022). There are many instances of algorithms designed to bring people together inadvertently driving separation (e.g. dating apps, social media, HR tools, loan approvals)(Bivens & Hoque 2018; Garcia, Garcia & Rigobon 2024; Köchling & Wehner 2020; McDavid 2020; Nader 2020; Narr 2021). There is a growing understanding, perhaps incited by fear of harsh legislation being introduced, that rectifying issues as they occur is inadequate and safe guards must be implemented to prevent future harm (Aysolmaz, Iren & Dau 2020; Rudin 2019; Schäfer & Wiese 2022).One way to prevent this harm is to prevent algorithms from detecting a bias. The surest method is obviously to remove the bias in reality.

If algorithms can learn these ‘patterns of bias’, then it stands to reason that they themselves could be used as tools of detection for that same pattern. This idea gave birth to the questions:

> 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?

There are simple frameworks that hold merit, such as the implicit association test, implicit relational assessment procedure, Go/No-Go association task, self-reporting measures, or HR complaints policies (Citro, Dabady & Blank 2004; Greenwald et al. 2022; Smith et al. 2015). However, expecting individuals to partake accurately in a survey on their own implicit bias is unrealistic and dangerous. The above methods focus on the ‘perpetrator’, who may not be cognisant of the bias, incapable of reporting on it, or inclined to present themselves in a more aggregable light (Kang & Lane 2010; Lugon Arantes 2021). There is also no legislation

mandating that these activities take place, and the time, cost and eventual reform they require do not make them an attractive option. However, the technique suggested here is low cost, uses pre-recorded datasets that could be reasonably expected to have latent information regarding bias (and so not impacting worker productivity), and is considerably faster. It also has the advantage of being simple to explain and computationally efficient, making it a practical tool for organizations to proactively address bias.

To demonstrate and measure the technique, an exemplar dataset has been selected from the *UC Irvine Machine Learning Repository.* The data was initially compiled for the paper titled '*Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records'* published in *BioMed Research International (Strack et al. 2014)*. It has 101,766 instances with 47 features (identity-related and treatment/experience-related features) and contains data collected from 1999-2008 from 130 hospitals across America. The data has been deidentified and has the appropriate licensing to be used for this purpose. The American medical system was selected because it has a well-documented history of bias (Feagin & Bennefield 2014; Galvan & Payne 2024; Jindal et al. 2023). The methods trialled during this process included K-means, density based spatial clustering applications with noise (**DBSCAN**), and hierarchical density based spatial clustering applications with noise (**HDBSCAN**). In the end the most computationally efficient model with the most meaningful output was a uniform manifold approximation and projection (**UMAP**) and HDBSCAN combination.

Literature review

Clustering models use mathematical techniques to reveal groups of similar instances in datasets (Ezugwu et al. 2021). An early technique was hierarchical clustering, a nested clustering approach using distances between instances themselves to determine cluster membership (Nagpal, Jatain & Gaur 2013). While hierarchical clustering provides a dendrogram with explanatory power, it is a very simple model. Another early model, the K-means algorithm, measures instance-to-centroid distances to determine cluster membership, however it is sensitive to noise and non-spherical cluster shapes (Bock 2007). The broad utility of clustering models created the desire for application on more complex datasets. The 1990's introduced the technology that made this desire computationally feasible (Lim 2019). With increased computing capacity, newer clustering algorithms were conceived, no longer measuring instance's -to-centroid proximities. Instead, new density-based algorithms put labels to pockets of density in the dataspace (Nagpal, Jatain & Gaur 2013). Models like DBSCAN were insensitive to noise and enabled non-spherical cluster shape identification (Nagpal, Jatain & Gaur 2013). More modern techniques include expectation-maximization algorithm, which relied on an instance's feature distributions, using probability to determine cluster membership, and providing a flexible means of clustering complex data (Nagpal, Jatain & Gaur 2013). However, more complex models tend to require extensive domain specific knowledge to inform hyperparameter choices and provide little in the way of explanation (Vázquez, Zseby & Zimek 2020; Yang, Jiao & Pan). One problem that remains is the scalability of these models, particularly with the immense amount of data resulting from the fourth industrial revolution (Drobot 2020). Parallelisation techniques assisted to some degree, but more computationally efficient alternatives are being conceived.

A team of researchers at the *Tutte Institute for Mathematics and Computing* developed HDBSCAN on python (McInnes, L, Healy & Astels 2017). This model uses mutual reachability distances between instances to imitate a probability density function that is then used to identify clusters. It allows for decomposition of clusters into smaller cluster, pruning off decompositions below the selected 'min_points' threshold. To further the efficiency of this method, McInnes et al. constructed UMAP as a means of reducing dimensionality in a manner that conserves local structures and separates them from unrelated structures (2018). This technique is mathematically complex, but essentially it creates a graph-like complex that approximates the relationships in the data well, but in fewer dimensions (see appendix for details). It outperforms alternative dimensionality techniques (eg, T-SNE), excelling with large datasets (McInnes, LH, John; Melville, James 2018).

Clustering models as a tool for bias identification is a relatively recent concept. After lengthy review, just one paper could be found which clustered biased language to determine how this bias may influence natural language processing applications (Caliskan et al. 2022). The discussion is predominantly regarding clustering for segmentation analyses or similar, and how this can result in biased outcomes (Klein 2016; Nakip, Gökmen & Mohammed 2017). Fewer still discuss the inevitable situation of patterns of bias being identified inadvertently, though researchers make no direct comments (Recchia et al. 2022). There does not appear to be any literature presenting clustering as an application to directly identify potential bias. Clustering algorithms can effectively draw on patterns that aren't necessarily observable or easily measured. This paper seeks to demonstrate their utility in this space, promoting the use of data science in corporate-social responsibility and providing a robust tool that is economic, efficient, and conceptually simple.

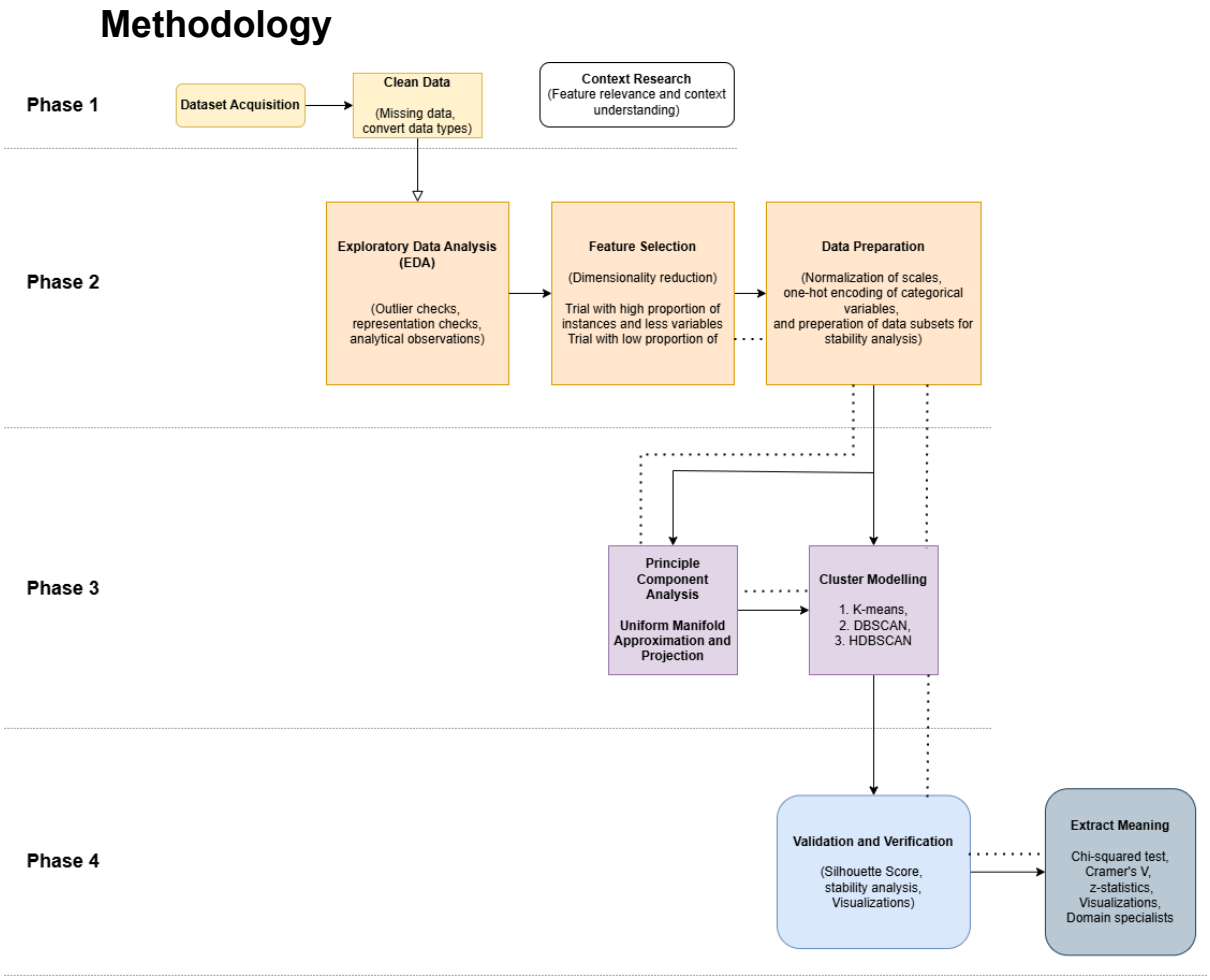

Figure 1: The method for testing clustering as a means of identifying potential patterns of bias.

Cleaning and Preprocessing

Python was used to separate identity features which remained unknown to prevent researcher bias. Four datasets were constructed from the original while the best instance-to-feature ratio was determined. They included:

- High instances set – 99493 instances with 16 features
- Low instances - 27140 instances with 18 features
- Principal components – 99493 with 14 principal components
- UMAP components – 99493 with 2 components

*Table 1: Feature selection explanations. *Features marked were only included in low instance dataset.*

Feature	Reason for inclusion
Admission type	Patterns of admission are contextually important and may indicate attitudes (Wild et al. 2010)..
Discharge disposition	The way in which a patient was discharged is indicative care received and/or complexity (Spooner et al. 2017).
Time in hospital	The number of days between admission and discharge. Indicative of cost incurred and received care (American Diabetes Association 2018; Kapoor et al. 2011; Niohuru 2023).
Medical speciality*	Context of the doctor and flags suboptimal care (Hashem, Chi & Friedman 2003; Singh & Venkataramani 2024).
Payer code*	Payment type (e.g. insurance, Medicare/Medicaid, out-of-pocket, etc.). Has socioeconomic implications (Kapoor et al. 2011).
Number of laboratory procedures	Indicative of cost in assets hospital was willing to incur, affordability to patients, case complexity, or the degree of attention, focus, and commitment received from the medical team.
Number of non-laboratory procedures	May capture instances of inadequate care/ neglect and/or a poorly equipped hospital.
Number of medications received	Indicative of affordability and/or case complexity. May also capture patients receiving the wrong treatment from lack of attention or hospital resources.
Times as an inpatient in the last year	An indication of medical complexity, age, overall health, and socioeconomic status (Naik et al. 2024).
Number of diagnoses	Indicative of case complexity, a potential measure of attention and more generally of system failure (Fraser et al. 2010)
Change made to patient's medication plan	Indicates that a patient had prior medication that was not meeting their health needs, perhaps indicating a history of suboptimal care, worsening condition, more attentive clinicians during the secondary treatment while in hospital, or the treatment failure (Auerbach et al. 2016).
Readmission	If a patient returned to hospital. Potentially due to negligent or suboptimal care received initially (Auerbach et al. 2016).

Table 2: Identity features required to identify potential bias

Identity Feature	Reason for inclusion
Age	10-year intervals of age
Ethnicity	Ethnicity recorded (unclear if reported by hospital or patient)
Gender	Gender recorded (unclear if reported by hospital or patient)

All numerical features were discrete and normalized with MinMax scaling because distributions were skewed and contained several outliers. Normalizing was necessary to preserve distribution shapes on comparable scales. Categorical features were one-hot encoded, resulting in 16 features expanding to 45. Principal component analysis, **PCA**, was conducted to observe feature importance and reduce dimensions down to 14 components (capturing 95% of the variance). UMAP was also performed on the high instance subset to

reduce the dimensionality down to only 2 components, a reverse transformation to measure the error in the reconstructed dataset was used to determine information loss.

Model Selection

Each subset was split into 5 folds that could be used to determine cluster stability with silhouette scores or centroid locations. Silhouette scores were selected because they don't require a ground truth. K-means models were performed on each subset over a range of K values. A grid search was attempted with DBSCAN to obtain the best model, but this was too computationally expensive. An UMAP + DBSCAN combination, was used in conjunction with a grid search (across min_dist, n_components, and n-neighbours for UMAP and min_samples and minimum_cluster_size for the HDBSCAN) on five folds of the data to obtain the best model as measured by silhouette scores.

Statistical Analysis

The identity features (separated from the beginning so neither researcher nor the model had access to them) were then used to determine patient populations in each cluster. A chi-squared test of independence, Cramer's V, and z-tests were used to identify statistical significance (p-value ≤ 0.05).

Note – This method is not designed to prove a bias, it is designed to identify potential patterns of bias, requiring domain specific professionals to engage with the findings.

Ethical Considerations

Reidentification

Patient data could theoretically be used to reidentify patients. However, this would require additional information protected by American medical institutions and privacy laws.

The Quality of Representation

Only people that trust the system with their care, are granted access, or can afford medication are represented. This implies that not all patterns of bias are represented in this data. That is not to say there are no patterns present. In practice, more datasets would be required to ascertain the full picture. Affordability of care is a societal form of bias and will be treated as such. Health care affordability is outside the scope of this research, but systemic inequality is acknowledged as another way in which the American medical system can be discriminatory (Jindal et al. 2023). Finally, data collection surveys often don't provide a way to record non-binary genders accurately, resulting in erasure. A form of bias that will be undetectable with this technique.

Correlation and Not Causation

The dataset may contain patterns of genetics, lifestyle, etc. Modelling may create the impression that patterns in the data must equate to a bias, though it is worth investigating, it is by no means 'proof' of bias. The findings will always require domain experts to assess the association.

Researcher Bias

Every precaution has been taken to ensure that clusters reflect only the patterns from the data and have not been coerced in any way. To prevent outside influence, the identity features were concealed until the final model was validated.

Results

Dimensionality reduction

PCA:

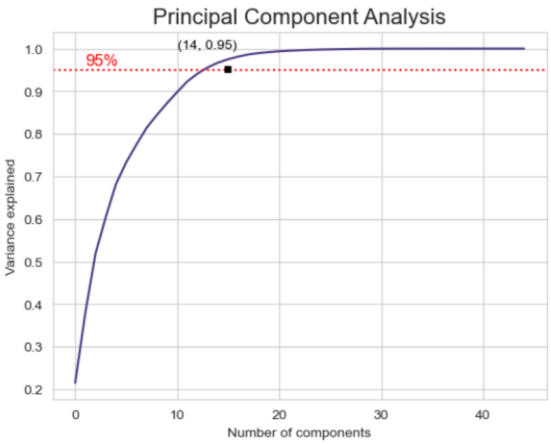

Figure 2: PCA cumulative plot of variance explained ratio.

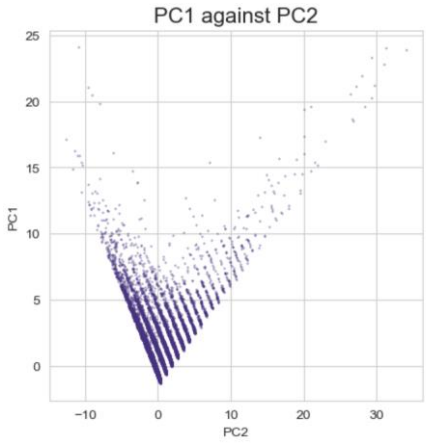

Figure 3: The two largest principal components mapped against each other.

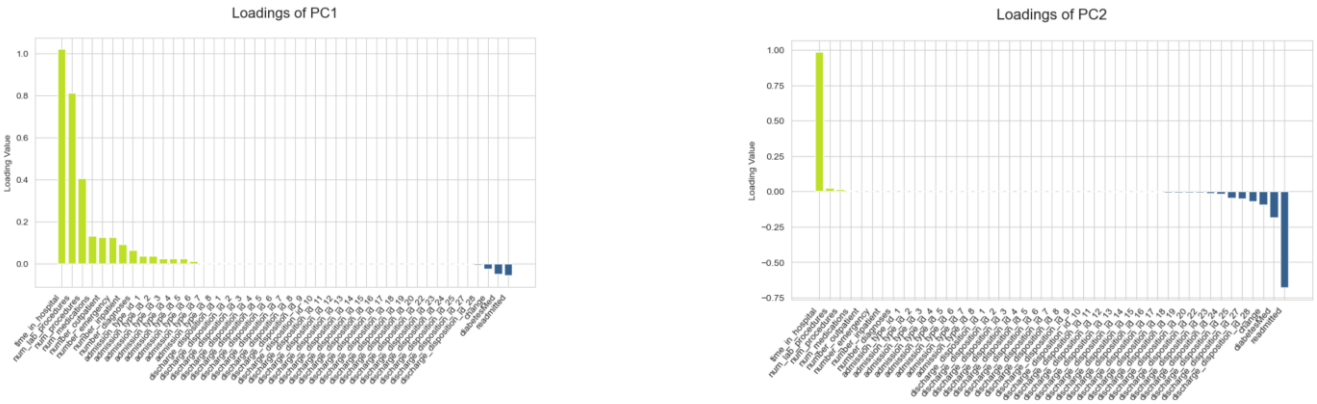

Figure 4: The loading scores observed for each feature in PC1 and PC2. These relate to the features importance in explaining variance.

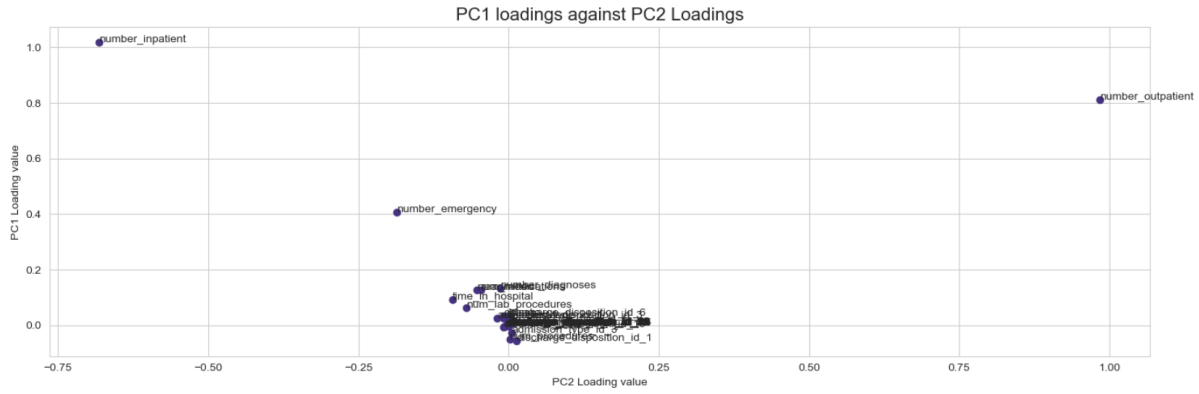

Figure 5: The loading plot reveals a cluster of one-hot encoded variables close to each other at the origin. The greater the distance between variables in the loading plot, the less correlated they are. Although several features are close to the origin, this does not necessarily imply there is no meaning in them.

Approximately 14 components could explain ≥95% of the variance (figure 2). It was not clear that clustering models would perceive patterns from this transformation as loadings didn't place great importance on several features (many being one-hot encoded). Plots of PC1 and PC2 was quite uniform (figure 3, 4 & 5).

UMAP:

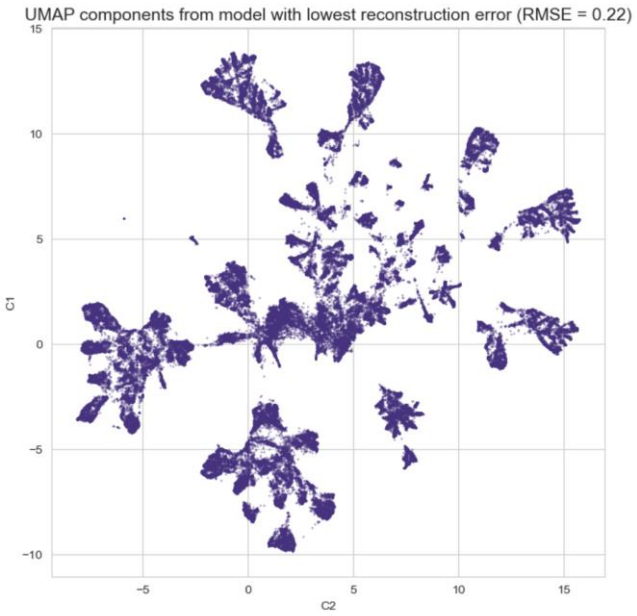

Figure 6: Embedding of the 2 component UMAP with the lowest reconstruction error (RMSE = 0.22).

The model with the lowest reconstruction error used 2 components, 15 neighbours, and a minimum distance of 0.05. As data was pre-processed to contain only numerical values between 0 and 1, the root mean squared error of 0.22 may appear high. This does indicate a loss of some information, however UMAP captures topological structures and does not intend to preserve exact values. An error of 0.22 across two transformations (there and back) was deemed acceptable given the abstract nature and variance of the dataspace with such a vast difference in dimensions.

Model selection

K-means:

K-means algorithms were applied to low instance, high instance, PCA and UMAP transformed datasets. The performance reduced as K increased for all except the UMAP embedded dataset. Likely due to the local structures better approximating isolated spheres, however the arbitrary choice of K remained a concern.

Table 3:Example output - K-means model on the high instance dataset.

	K	Fold 1 - Silhouette scores	Fold 2 - Silhouette scores	Fold 3 - Silhouette scores	Fold 4 - Silhouette scores	Fold 5 - Silhouette scores	Fold 1 - Distortion	Fold 2 - Distortion	Fold 3 - Distortion	Fold 4 - Distortion	Fold 5 - Distortion	Mean Silhouette score	Mean distortion
0	2	0.367208	0.369094	0.361912	0.366434	0.362439	628250.168263	627847.496241	617838.478071	619609.688583	619565.390272	0.365418	622622.244286
1	6	0.122277	0.103570	0.110444	0.083396	0.095510	457017.359170	456940.676120	456512.249205	459672.707284	456724.480837	0.103039	457373.494523
2	10	0.090903	0.097305	0.098461	0.098890	0.090226	389826.038226	386761.340540	375653.660005	375882.564934	385602.913079	0.095157	382745.303357
3	14	0.091363	0.088908	0.081892	0.086948	0.085743	352131.749990	342933.908572	348477.584402	343943.171802	349953.082094	0.086971	347487.899372
4	18	0.085774	0.077960	0.088412	0.083328	0.083034	318932.289514	318584.845312	322302.871446	320295.991033	320657.131307	0.083701	320154.625722

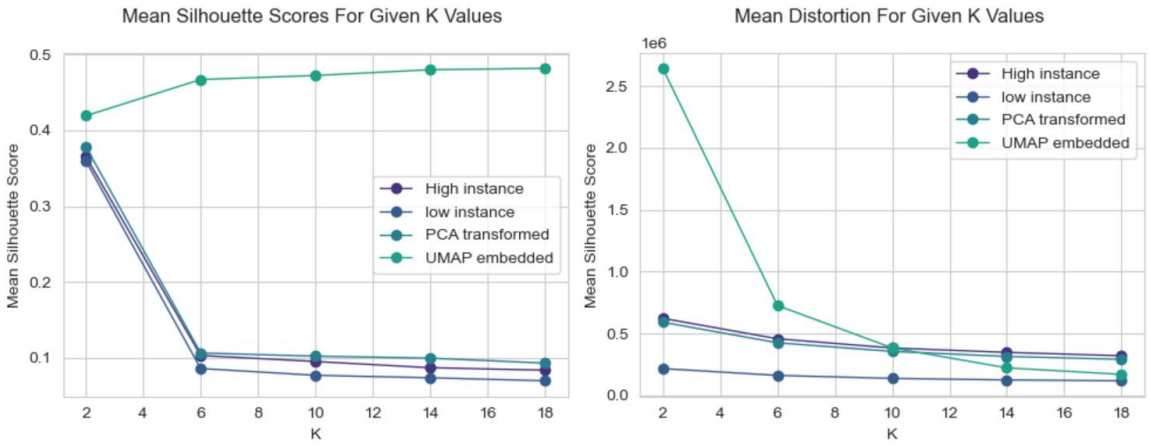

Figure 7: Mean silhouette score and mean distortion observed for each dataset for various k-values

DBSCAN:

This model was very computationally expensive, with the kernel failing 44 hours in. This inefficiency, and the reliance on epsilon, implied it was an impractical choice. This led investigations into UMAP and HDBSCAN.

HDBSCAN:

Processing times were initially unacceptable. As UMAP was intended explicitly to aid HDBSCAN, this combination was the only one trialled due to time constraints and excessive processing times. Processing times were considerably reduced with this pipeline. The tuning phase included 76 combinations of ‘min_cluster_size’ and ‘min_samples’ values. The model with the best silhouette score (mean train score of 0.333 and standard deviation of 0.006) used 85 as the minimum cluster size and 8 as the minimum samples size. This was deemed acceptable as the clusters were observably non-spherical and many neighboured each other quite closely, the score was therefore considerably high. Given the size and variance of this dataset, a degree of noise was anticipated. This model marked 15.31% of the data as noise while finding 280 clusters which offered a very granular view.

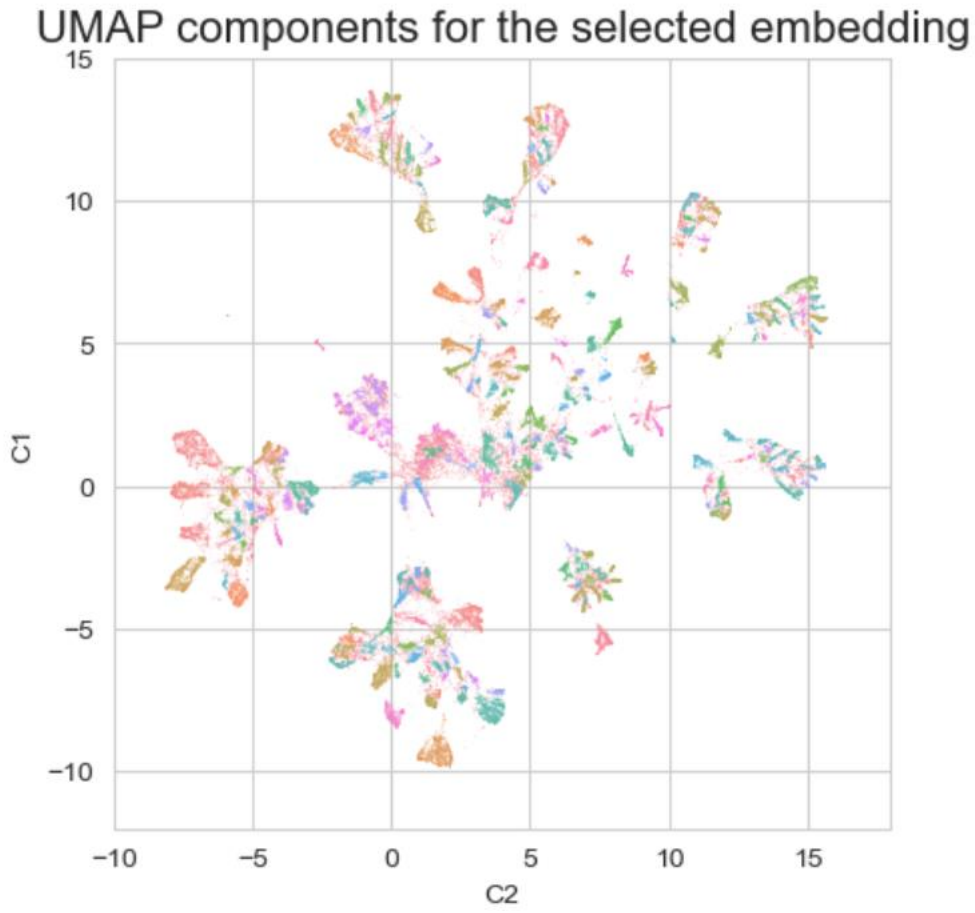

Figure 8: A visualization of the 280 clusters identified by the UMAP+HDBSCAN combination.

Figure 9: Three random selections to observe granularity and compactness of clustering.

The above was all conducted using modules in scikit-learn as it provided hyperparameter tuning tools and pipelines. However, the associated HDBSCAN method is quite limited, so an attempt was made using another module to find parameter sets that perhaps don't have the best silhouette score but offer a less granulated output.

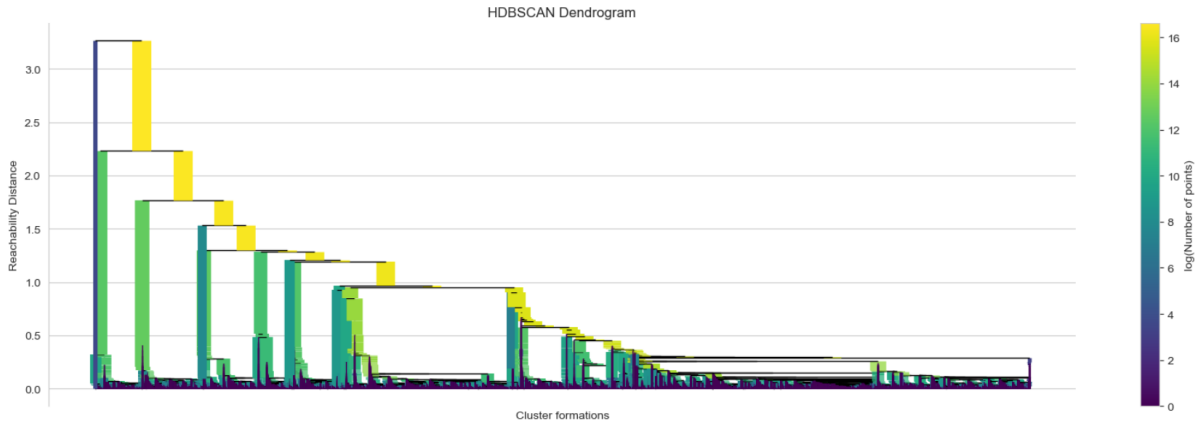

Figure 10: Dendrogram demonstrating the cluster formations output by HDBSCAN at various mutual reachability distances.

It became apparent that the two modules have variations that would require hyperparameters to be re-tuned for the new model (without the aid of scikit-learns pipelines). The time required and expected benefit prevented further investigation, though future researchers may benefit from having access to the dendrogram to determine the level of granularity desired. Additionally, this secondary HDBSCAN module provided access to mutual reachability distance which is the distance metric also used by UMAP. Future researchers may consider an alternative to the silhouette score that uses this distance metric as it might provide a more nuanced and meaningful measure of verification/validation.

Stability analysis

Minor changes were applied to the UMAP parameters to observe how this impacted the shape identified in 2 components. Alterations didn't appear to alter the shape of the plotted embedding, suggesting that the parameters selected were quite robust in this range. HDBSCAN can only attempt to find and label existing dense groups, and each fold applied in the cross validation will label them differently. However, a standard deviation of 0.006, and over 200 clusters found per fold (despite a fifth of the data being withheld) indicate there is similar Euclidean distance and density behaviours across the five folds. With a fifth of the data missing from each fold it is unsurprising that some clusters drop below 85 core points and are removed.

Statistical analysis

Identity features were reintroduced, and cluster proportions of ethnicity, gender, and age were investigated. It was observed that some of the clusters were considerably different from the population proportions. After removing the instances associated with noise, these discrepancies were measured as follows:

- Chi squared test of independence – To determine if an identity feature has a meaningful association with cluster labels.
- Cramer's v – To determine the weight of an association if found.
- Z-statistic – To measure the variance between a given group in a cluster and the population mean.

Table 4: The results obtained from the chi-squared test of independence across the three 'binarized' identity features.

Identity feature	Chi-squared test (p-value, degrees of freedom)	Cramer's V
Race (Caucasian/ non-Caucasian)	2915.44 (~0.0, 279)	0.186 Weak - Moderate
Gender (Male/ Female)	1386.43 (~0.0, 279)	0.128 Weak - Moderate
Age (60-68 / all others)	1394.87 (~0.0, 279)	0.129 Weak - Moderate

Three subgroups were selected for targeted investigation with z-statistics. This included African Americans, women, and those at retirement age (60-69).

African Americans – 91 clusters were found to have z-statistics indicating the proportion of African Americans was at least three standard deviations away from the expected population proportion (p-value ≤ 0.05). These clusters each contained between 0.102% and 1.984% of non-noise instances. Of these:

- The cluster with the biggest variation was cluster 266 (z-stat = 14.50)
- The cluster with the highest proportion of African Americans was cluster 27 (42.81%), and cluster 249 had the largest proportion of Caucasians (93.29%).
- The lowest proportion of Caucasians was 49.43% in cluster 24

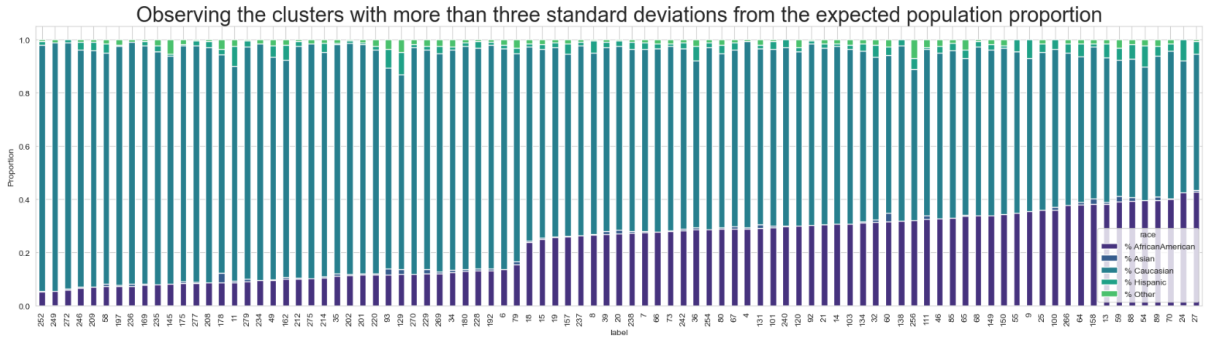

Figure 11: The clusters at least three standard deviations away from the expected proportion of African Americans.

Women – 41 clusters were found to have z-statistics indicating the proportion of women was at least three standard deviations away from the expected population proportion (p-value ≤ 0.05). These clusters each contained between 0.114% and 2.244% of non-noise instances. Of these:

- The cluster with the largest variation was cluster 180 (z-stat = 8.69)
- The cluster with the highest proportion of women was cluster 188 (73.75%)
- The cluster with the least proportion of women was cluster 105 (32.98%)

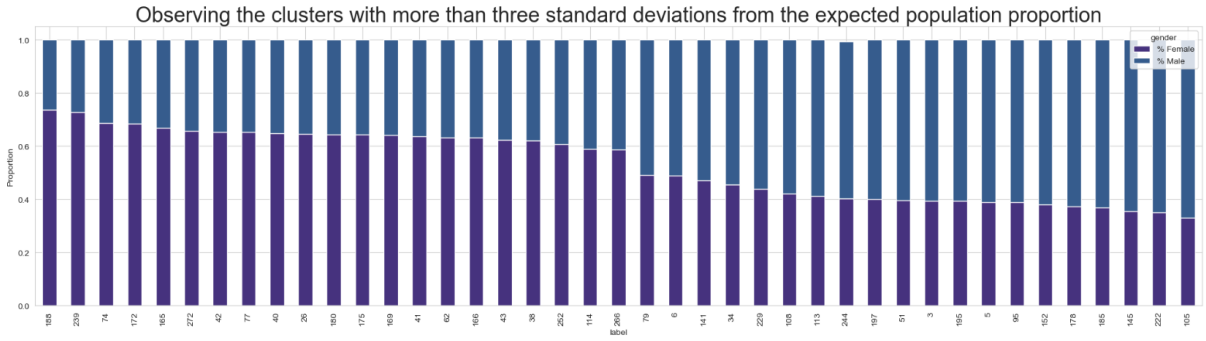

Figure 12: The clusters with at least three standard deviations away from the expected proportion of Women.

60–69 year olds – 42 clusters were found to have z-statistics indicating the proportion was at least three standard deviations away from the expected population proportion (p-value ≤ 0.05). These clusters each contained between 0.223% and 2.244% of non-noise instances. Of these clusters:

- The cluster with the largest variation was cluster 222 (z-stat = 6.52)
- The cluster with the highest proportion of people aged 60-69 was cluster 222 (38.27%)
- The cluster with the lowest proportion of people aged 60-69 was cluster 60 (4.97%)

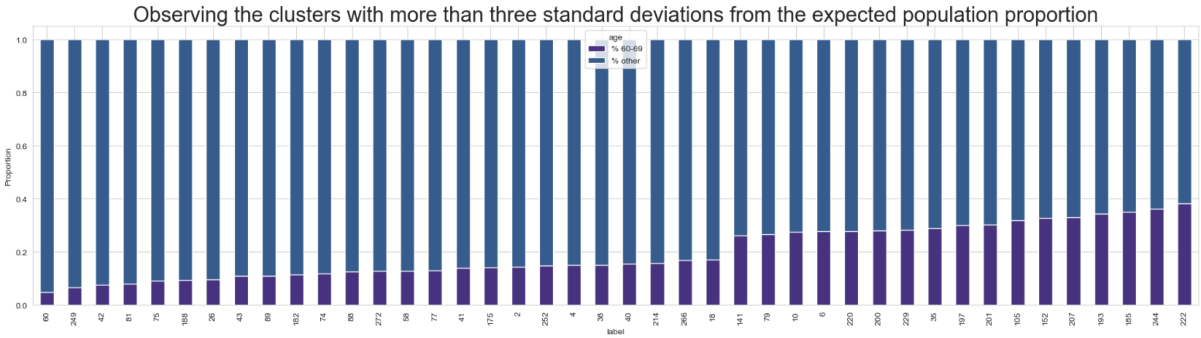

Figure 13: The clusters with at least three standard deviations away from the expected proportion of 60-69 year olds.

Discussion

Medical records were found to be incomplete, very few weight values were recorded which is disappointing as perceptions around weight are known to influence behaviour (Lawrence et al. 2021; Moerschel 2021). With patterns theorised to reside in latent information, and not the features themselves, four data subsets were constructed to find a balance between instances and information load. As the project evolved, the inclusion of additional features in the low instance set was reconsidered as their information load and added variance were unfavourable. The high instance sets were more comparable, and therefore more meaningful. One of the secondary aims of this project was to ensure computational accessibility. Dimensionality reduction was necessary to reduce processing requirements. PCA uses orthogonal perspectives to capture variance and reduce noise (Louhichi et al. 2023). The PCA captured over 95% of the variance in 14 dimensions, however this preprocessing step did not reveal obvious clusters visually in the first two components, potentially due to the discrete nature of all pre-processed variables (*figures 2 & 3*). The alternative method, UMAP, was more successful, making clusters observable in just two dimensions. UMAP was reverse transformed with a mean squared error of 0.05 (RMSE = 0.224). This is a reasonable reconstruction score (given reduction of 45 dimensions down to just two) and implies that the major patterns are well represented in the two-dimensional space.

The K-means models showcased that it is impractical for this purpose, as clusters are unlikely spherical or noiseless in any domain, two key requirements of the model (Bock 2007). While K-means seemed capable of clustering the UMAP embedded dataset, It took a great deal of time. Similarly, DBSCAN was found to be computationally inefficient making it impractical for common use in this manner. However, UMAP and HDBSCAN were able to capture local patterns and group them in a way that could be explained simply, visualized clearly, modified for purpose, with sufficient verification and validation checkpoints, and had the advantage of handling varying densities well. Structures identified in the UMAP were found to be stable across a range of parameter values. The HDBSCAN found similar distance behaviours and areas of density across the five folds during the cross-validation. Confidence in UMAP was earned through its previous measurable success in supervised settings (Lawrence et al. 2021). Future research should seek to strengthen verification methods around the embedding of non-supervised data as this underpins the accuracy of the entire approach.

This model has taken pre-recorded data with features reasonably expected to capture patterns of bias within their latent information and found patterns in medical experiences for a population of patients. This data is generally difficult to tamper with and not anticipated to be used for this purpose, offering some protection from obfuscation. Blame isn't place on an individual, instead discussions are directed towards the medical experience and the processes that enabled that experience. The output from the model provides researchers with groups of interest, directing attention towards groups with experiences that diverge most from the generic treatment. However, it must be noted that not all groups have the same density or membership. A domain specialist is needed to assess these clusters and determine their medical validity. This specific dataset includes a decade of experiences from 130 different hospitals, it is unsurprising that several clusters were identified yet somewhat surprising that no national trend was identified given the aforementioned literature regarding the well documented bias of the American medical system. Future researchers should attempt to only use the trailing 6 months for a given region. This factor may have obfuscated the findings, for example a once common experience may now be suboptimal and only offered to patients of

low socio-economic status or victims of bias; this may then result in a cluster with older-privileged and newer-underprivileged patients muddled together by time. Another confounding factor is that a given ethnic group may be concentrated around an underfunded hospital resulting in unfavourable associations. However, if something has the appearance of bias, then attention is needed whether or not bias is present (unless there is a genetic or cultural component to justify the difference).

Race, gender, and age were all found to be weakly associated with cluster labels. Many of the clusters were significantly different from the expected proportion of a given identity feature. Domain specialists should start with the clusters that contain the most variance, determining the validity of the behaviours observed in that experience (clusters 180, 222, 266). Ideally the domain specialists remain blind to the identity features associated, only being prompted after their assessment if it would benefit a person of a given race, age, or gender to avoid their internal biases from becoming a factor. There may be patterns of bias identified that require further investigation, however, these investigations may simply highlight the need for policy changes, retraining, or improved funding in hopes of achieving better outcomes for all patients and not just the group perceived to be disadvantaged.

Conclusion

Although the dataset had its shortcomings, the project has demonstrated a methodology that appears to provide domain specialists with information required to identify potential bias, and to locate the impacted groups in a population. It does so at low computational and organizational cost in a non-accusatory manner that encourages open dialogue, isn't reliant on self-reporting or survey responses, and has the advantage of being conceptually simple. While other clustering models struggled, the UMAP+HDBSCAN combination has the capacity to assist with the identification of patterns of bias in pre-recorded data in a broadly accessible manner. Future research should seek to apply these tools on other datasets to confirm its utility. There is a growing need for real-time modelling to flag potential bias and provide decision makers an opportunity to reflect on their decision before serious harm occurs.

References

Auerbach, AD, Kripalani, S, Vasilevskis, EE, Sehgal, N, Lindenauer, PK, Metlay, JP, Fletcher, G, Ruhnke, GW, Flanders, SA, Kim, C, Williams, MV, Thomas, L, Giang, V, Herzig, SJ, Patel, K, Boscardin, WJ, Robinson, EJ & Schnipper, JL 2016, 'Preventability and Causes of Readmissions in a National Cohort of General Medicine Patients', *JAMA internal medicine*, vol. 176, no. 4, pp. 484-493.

Aysolmaz, B, Iren, D & Dau, N 2020, 'Preventing algorithmic Bias in the development of algorithmic decision-making systems: A Delphi study'.

Bivens, R & Hoque, AS 2018, 'Programming sex, gender, and sexuality: Infrastructural failures in the “feminist” dating app Bumble', *Canadian Journal of Communication*, vol. 43, no. 3, pp. 441-459.

Bock, H-H 2007, 'Clustering Methods: A History of k-Means Algorithms', in Springer Berlin Heidelberg, pp. 161-172.

Caliskan, A, Ajay, PP, Charlesworth, T, Wolfe, R & Banaji, MR 2022, 'Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics', paper presented at Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom.

Citro, CF, Dabady, M & Blank, RM 2004, *Measuring Racial Discrimination*, National Academies Press, Washington, D.C., UNITED STATES.

Drobot, AT 2020, 'Industrial Transformation and the Digital Revolution: A Focus on Artificial Intelligence, Data Science and Data Engineering', in IEEE.

Ezugwu, AE, Shukla, AK, Agbaje, MB, Oyelade, ON, José-García, A & Agushaka, JO 2021, 'Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature', *Neural Computing and Applications*, vol. 33, no. 11, 2021/06/01, pp. 6247-6306.

Fraser, L-A, Twombly, J, Zhu, M, Long, Q, Hanfelt, JJ, Narayan, KMV, Wilson, PWF & Phillips, LS 2010, 'Delay in Diagnosis of Diabetes Is Not the Patient's Fault', *Diabetes Care*, vol. 33, no. 1, pp. e10-e10.

Garcia, ACB, Garcia, MGP & Rigobon, R 2024, 'Algorithmic discrimination in the credit domain: what do we know about it?', *AI & SOCIETY*, vol. 39, no. 4, pp. 2059-2098.

Greenwald, AG, Dasgupta, N, Dovidio, JF, Kang, J, Moss-Racusin, CA & Teachman, BA 2022, 'Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem', *Psychological Science in the Public Interest*, vol. 23, no. 1, pp. 7-40.

Jindal, M, Chaiyachati, KH, Fung, V, Manson, SM & Mortensen, K 2023, 'Eliminating health care inequities through strengthening access to care', *Health Services Research*, vol. 58, no. S3, pp. 300-310.

Kang, J & Lane, K 2010, 'Seeing through colorblindness: Implicit bias and the law', *UCLa L. rev.*, vol. 58, p. 465.

Klein, TA 2016, 'Exploring the ethical and societal implications of market segmentation and targeting: Macromarketing and distributive justice perspectives', *Social Business*, vol. 6, no. 2, pp. 109-124.

Köchling, A & Wehner, MC 2020, 'Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development', *Business Research*, vol. 13, no. 3, pp. 795-848.

Lawrence, BJ, Kerr, D, Pollard, CM, Theophilus, M, Alexander, E, Haywood, D & O'Connor, M 2021, 'Weight bias among health care professionals: a systematic review and meta-analysis', *Obesity*, vol. 29, no. 11, pp. 1802-1812.

Lim, TW 2019, *Industrial revolution 4.0, tech giants, and digitized societies*, Springer.

Louhichi, M, Nesmaoui, R, Mbarek, M & Lazaar, M 2023, 'Shapley values for explaining the black box nature of machine learning model clustering', *Procedia Computer Science*, vol. 220, pp. 806-811.

Lugon Arantes, PDT 2021, 'The Due Diligence Standard and the Prevention of Racism and Discrimination', *Netherlands International Law Review*, vol. 68, no. 3, pp. 407-431.

McDavid, J 2020, 'The social dilemma', *Journal of Religion and Film*, vol. 24, no. 1, pp. 0_1-3.

McInnes, L & Healy, J 2018, 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', 02/09.

McInnes, L, Healy, J & Astels, S 2017, 'hdbscan: Hierarchical density based clustering', *The Journal of Open Source Software*, vol. 2, no. 11, p. 205.

McInnes, LH, John; Melville, James 2018, 'UMAP: Uniform Manifold Approximation and Projection Documentation', viewed November 15, 2024, <https://umap-learn.readthedocs.io/en/latest/index.html>.

Moerschel, L 2021, 'The Intersectionality of Anti-fat Prejudice'.

Nader, K 2020, 'DATING THROUGH THE FILTERS', *Social Philosophy and Policy*, vol. 37, no. 2, pp. 237-248.

Nagpal, A, Jatain, A & Gaur, D 2013, 'Review based on data clustering algorithms', in *2013 IEEE Conference on Information & Communication Technologies,* pp. 298-303.

Naik, H, Murray, TM, Khan, M, Daly-Grafstein, D, Liu, G, Kassen, BO, Onrot, J, Sutherland, JM & Staples, JA 2024, 'Population-Based Trends in Complexity of Hospital Inpatients', *JAMA internal medicine*, vol. 184, no. 2, pp. 183-192.

Nakip, M, Gökmen, A & Mohammed, SA 2017, 'Financial Market Segmentation: An Application on Islamic Financial Markets', *Journal of Applied Economics & Business Research*, vol. 7, no. 4.

Narr, G 2021, 'The Uncanny Swipe Drive: The Return of a Racist Mode of Algorithmic Thought on Dating Apps', *Studies in Gender and Sexuality*, vol. 22, no. 3, 2021/07/03, pp. 219-236.

Recchia, DR, Cramer, H, Wardle, J, Lee, DJ, Ostermann, T & Lauche, R 2022, 'Profiles and predictors of healthcare utilization: using a cluster-analytic approach to identify typical users across conventional, allied and complementary medicine, and self-care', *BMC Health Services Research*, vol. 22, no. 1.

Rudin, C 2019, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, vol. 1, no. 5, 2019/05/01, pp. 206-215.

Schäfer, J & Wiese, L 2022, 'Clustering-Based Subgroup Detection for Automated Fairness Analysis', in Springer International Publishing, Cham, pp. 45-55.

Smith, CT, Ratliff, KA, Ortner, T & Vijver, F 2015, 'Implicit measures of attitudes', *Behavior based assessment in psychology: Going beyond self-report in the personality, affective, motivation, and social domains*, pp. 113-132.

Strack, B, Deshazo, JP, Gennings, C, Olmo, JL, Ventura, S, Cios, KJ & Clore, JN 2014, 'Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records', *BioMed Research International*, vol. 2014, pp. 1-11.

van Giffen, B, Herhausen, D & Fahse, T 2022, 'Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods', *Journal of Business Research*, vol. 144, 2022/05/01/, pp. 93-106.

Vázquez, FI, Zseby, T & Zimek, A 2020, 'Interpretability and Refinement of Clustering', in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA),* pp. 21-29.

Yang, H, Jiao, L & Pan, Q 'A Survey on Interpretable Clustering', in IEEE.

Appendix

Link to dataset:

https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

UMAP Explanation:

Abstract simplicial complexes (abstract shapes that generalise space) and nerve theorem (a nerve and the union of the sets have the same homotopy type allowing simplicial complexes to connect and create topological spaces) are used to effectively capture local structures/information in a compressed way. This topological space is unlikely to have a perfect distribution, so usual distance metrics cannot be used to create a single topological space. To connect these complexes into a single topological space, the assumption that they are normally distributed but on a curved surface is forced, thus creating Riemannian distance metric. With this, all nerves and topological spaces will fuse along this 'manifold'. Then these finite spaces are converted to fuzzy simplicial sets allowing proximity to correlate with confidence of connection between sets. Cross entropy is uses to optimize the shape of the manifold, seeking to get the local units grouped and to get the global separation as accurate as it can. The result is that pockets of density in the high dimensional space get separated in a low dimensional space.

HDBSCAN Explanation:

The model calculates core distances between every point, it then calculates the reachability distance between pairs of points. It then constructs a minimum spanning tree and condenses using the reachability distances to remove edges of low density. Points outside of the dense regions are considered noise. As the density threshold changes, branches of tree are pruned off leaving stable branches which represent clusters.

Silhouette Score Explanation:

The silhouette score can be used in non-supervised learning techniques as it uses distance and isn't reliant on true labels. It indicates how well a cluster is matched to its own cluster compared to other clusters, with scores ranging from -1 to 1 (poorly matched to precisely matched). Mathematically, it's the difference between the average distances from a point all other points in that cluster and the average distance to all points from another cluster divided by the maximum of these two values. Balancing measured of cohesion and separation within the one metric.