Literature review: Clustering as a Means of Detecting Discrimination

Introduction

Machine learning continues to become a prominent part of everyday life. It is perhaps as prominent an innovation as the introduction of computers themselves. Both have exponentially improved productivity and enabled significant change. However, these tools are wielded by human hands in systems with pre-existing issues. When these tools were introduced, complete agreement had not been reached on key societal issues, and that remains true to some degree today. However, this effectively meant that the same stance on societal issues, and some of the context from prior to the tool's inception, are occasionally being implemented with these new tools in this new context (Miller, Peek & Parker 2020). This is true of issues such as racism, socioeconomic stratification, and gender identity or gender discrimination. The behaviour observed to date has been to advance and innovate quickly. In these modern times, with such powerful tools, there is ethical, financial, and political cause to slow down and take time to assess that these tools are being wielded in ways that build the society future generations deserve, and not simply regurgitating the implicit or explicit biases that historic discrimination introduced into key systems.

Historically, in-built discrimination or discriminatory cultures have been identified through audits or tests of system components, analysis of outcomes for specific cohorts, surveys, or complaints (Citro, Dabady & Blank 2004). Tests and audits are worthwhile, but they cannot create a clear picture of the whole system, nor does it necessarily provide an accurate measure of discrimination present in that particular part of the system (Adams et al. 1999; Smith et al. 2015). Many of these 'modern' methods rely on self-reporting, but expecting someone to report on their own implicit bias is unrealistic and dangerous. The individual may not be cognisant of the bias and potentially incapable of reporting on it or inclined to present evidence of unfair outcomes in a more aggregable light. A few of the more common tests include:

- Implicit association test (IAT) – asking individuals to categorize words or images into groups associated with gender, race, etc. It measures the strength of automatic pairwise associations (Greenwald et al. 2022; Smith et al. 2015).
- Implicit Relational Assessment Procedure (IRAP) – Similar to the IRAP, only instead of standalone words or pictures it asks the participants to what degree they agree with a particular phrase in a particular context. It measures the relationship between a concept and a group through relational framing, providing context dependent biases .(Smith et al. 2015)
- Go/No-Go Association Task (GNAT) – Similar to the IAT and IRAP. It measures a single category to determine implicit bias in that area (Smith et al. 2015).

These tests though imperfect, hold merit, but they are not mandated and so they will not drive change in organisations that don't prioritize or accept that injustice is occurring (Kang & Lane 2010; Lugon Arantes 2021). This author has proposed that clustering algorithms may be able to identify these behaviours in historical data and offer themselves as a means of providing some tangible evidence of bias that can incite discussions and ultimately make changes to outdated or failing systems in a way that is logical and non-threatening to unknowing perpetrators. This review will dissect this approach and present a case study candidate.

Machine Learning and Discrimination

Over the course of the last few decades, data scientists have built tools to extract meaning from data. Occasionally these tools have been constructed on bias or false assumptions, enabled biases, or learned a bias over time (van Giffen, Herhausen & Fahse 2022). For example, dating apps rely on algorithms to determine which two people have a high probability of finding chemistry. However, over time the algorithm learns biases from the users and promotes the separation of different groups preventing inclusion and engagement (Bivens & Hoque 2018; Nader 2020; Narr 2021). Another example was presented in the 2020 documentary titled '*The Social Dilemma'*, it explained how algorithms can result in polarization as users are provided with content similar to that of which they engaged with strongly prior and effectively solidifies an individual's world view (McDavid 2020). With growing concern in the scientific community and more broadly in the public domain, efforts continue to strengthen legislated protections (Schäfer & Wiese 2022). With the weight of public opinion and the threat of legislation, it is unsurprising that work has begun to correct inbuilt inequality within algorithms (Schäfer & Wiese 2022). It has been said that many attempts to correct bias occur after an injustice has occurred (Aysolmaz, Iren & Dau 2020). This highlights how the more complex the algorithm becomes, the less transparent its inner workings appear; thus, hindering the capacity for professionals to prevent harm. While the conversation begins to move towards preventing harm from occurring from artificial systems in the first place, the problem persists in the real world where these biases were first presented. These algorithms are tools, and they can be used to help remedy the problem they themselves face. Classification models can help us analyse population treatment, or mistreatment. Clustering models could theoretically be used to find patterns showcasing how different groups receive different treatment and provide information on how the treatment is different.

Clustering models

Clustering models operate on the principle that instances with similar attributes are more alike than those with differing attributes. They can uncover patterns using mathematical techniques to determine groups of similar instances (or clusters) in datasets, presenting a classification technique (Ezugwu et al. 2021). One of the first clustering techniques was hierarchical clustering, a nested clustering approach which uses the distances between instances themselves to determine what cluster they belong to (Nagpal, Jatain & Gaur 2013). Hierarchical clustering has the advantage of providing a dendrogram, a visualization of the nested clusters created throughout the process. Another early clustering model was a k-means clustering algorithm precursor, known then as Lloyd's algorithm, which was used as a classification tool in biology where an instance's distance from a given centroid (which can be thought of as a cluster "rally point") would determine to which cluster a given instance belonged (Bock 2007).

Advancements continued in both hierarchical and non-hierarchical clustering methods as their utility became more apparent and were applied more broadly in other fields. By the 1990s, computing power allowed for larger more complex datasets to be processed in a timely manner to great effect (Lim 2019). During this period, density-based algorithms were conceived and provided a new means of clustering such data, no longer relying on an instance's position in the 'dataspace' in relation to a centroid, but instead finding pockets of density in the dataspace (Nagpal, Jatain & Gaur 2013). Methods like Density-based spatial clustering of applications with noise (DBSCAN) enabled non-uniform cluster shapes to be identified and tolerated noise in ways that simpler models could not (Nagpal, Jatain & Gaur 2013). Other techniques were also introduced such as the expectation-maximization algorithm

(EM) which focussed on an instance's position on each of its features' distributions and then used probability to assign each instance to a cluster, providing another means of flexibly handling complex datasets (Nagpal, Jatain & Gaur 2013). However, as is common in data science, the more complex the model the less comprehensible the mechanics and interpretable the outcome. Simpler hierarchical models are unsupervised and provide a meaningful dendrogram to explain the outcome and decision process. More complex models may require some understanding of the expected number of clusters and provide little explanation for why an instance belonged to a specific group (Vázquez, Zseby & Zimek 2020; Yang, Jiao & Pan).

Eventually scalability of these models became an issue, particularly with the adoption of the internet and the immense amount of data that came along with it (Drobot 2020). While parallelisation techniques and more sophisticated models were conceived to handle this, they are beyond the scope of this report.

An Achievable Target

Machine learning models extract meaning from data and provide fantastic output provided the data scientist is skilled and uses clean and representative data. It stands to reason that these two requirements, difficult as they may be, will result in the desired utopia. To expand on thses requirements:

1. Requirement 1 – Professionalism
   Professional development teaching data scientists specifically that a model's verification must extend beyond mathematical measures of error and include an examination of its potential to cause harm and its ability to be influenced by external bias.
2. Requirement 2 – Managing Perception
   We must consider how we can promote equality to ensure our representative datasets are representing the world we want our models to be predicting in. That is, to build models using datasets that do not have discrimination or injustice represented within them.

Education is always a useful tool, but the more efficient is the latter. If society can rid itself of these biases, the ability for a model to do harm is greatly reduced. This is no simple task, but if machine learning models are creating inequality by detecting inequalities, then it is reasonable to expect they have the capacity to act as indicators of that same bias. In writing this review, there is currently little, if any, literature expressing how a machine learning model could be used as an indicator of bias in this way. However, machine learning could over tools that identify injustice disguised in the fabric of society and it's systems.

A Suggested Technique

Ridding society of all bias may be unachievable. Much of it is born of fear of the unknown which will always be part of the human condition. However, we have an obligation and a responsibility to ensure all living things are treated with dignity and respect. Many forms of injustice are born of a bias, be it explicit or implicit. With archaic systems, fear of judgement, and conservative tendencies toward the known, it is obvious why so many organizations have not made an intentional effort to seek out and treat the symptoms of bias or indeed cure the system. This author has suggested that clustering algorithms may be able to identify patterns of bias in historic data that encapsulates unmodified and representative behaviours. This would not indicate bias, but it may provide enough evidence (beyond the typical population statistics) to justify an organization tacking action without directly targeting any specific individuals. It may also present other social, financial, or legal arguments that are more persuasive to a governing body. To illustrate the concept, a medical dataset with information regarding diabetic patients will be used to assess the American medical system for bias.

Context on the American medical system

The American medical system must abide by both federal and state laws and is managed by the US department of Health and Human Services (Béland, Rocco & Waddan 2023). There is no universal healthcare coverage, but it is not uncommon for employers to provide health insurance in exchange for a reduced pay package (Arnold & Whaley 2020). Additionally, seniors or those living below the poverty line receive assistance from Medicare and Medicaid (Hoffman, Klees & Curtis 2000). Many hospitals also rely heavily on religious organizations to provide care (Bai, Yehia & Anderson 2020).

In America, insurance providers negotiate with specific doctors that they have pre-existing relationships with which perhaps creates a less dynamic system, but theoretically benefits patients with reduced expenses. In reality, these patients are typically charged considerably more than a patient in the Medicare or Medicaid public health system (Selden 2020). Barack Obama’s administration created legislation preventing insurers from refusing coverage because of pre-existing conditions, and made it mandatory for all persons without coverage through other means to obtain private health insurance (Béland, Rocco & Waddan 2023). There is assistance available to both find affordable cover and subsidies available to those who need them (Béland, Rocco & Waddan 2023). The government does not provide a cap on procedure costs as other nations with socialized healthcare systems do (Brown 2003).

The American health care system has long been viewed as capitalist system (Caplan 1989). However, consumers often aren’t positioned to look for the best deal when they need the care, which prevents capitalism from obtaining the best service at the best price. The healthcare system brings in considerable profits and holds powerful connections. There is significant resistance when changes are sought to make healthcare more affordable (Chua 2006; Gale 2019). A transition to a single-payer, or socialized, system similar to other nations may sound like a possible solution but this might initially have negative consequences for patients and doctors as they would need to see more patients in the same time period (less time per patient) to earn the same amount, hospitals that are more costly to run may even close, a consequence of running hospitals like a private business (Chua 2006; Emanuel 2008).

This system is also placed within a culture that struggles with racism, misogyny, and religious discrimination to various degrees in various states (Feagin & Bennefield 2014). This impacts the way individual doctors, staff, management, and the system interact and treat patients throughout America. This was made most apparent during the COVID-19 pandemic as reported in a recent Harvard public health video discussing these issues in-depth (Health 2021).

Specific considerations

It is understood that unwell patients may not approach the medical system due to fear of discrimination or because of the cost (Kannan & Veazie 2014). There are many documented patterns of bias in the American medical system that explain these behaviours (Feagin & Bennefield 2014; Galvan & Payne 2024; Jindal et al. 2023). While devastating, it does provide the perfect context to investigate this technique. For if this pattern can be identified in historic data unrelated to bias specifically and analysed in this context, then perhaps this technique can be applied to other contexts.

To avoid researcher bias, a clustering model will be constructed using features that may contain biased patterns, but not features associated with characteristics of the person, this includes their age, ethnicity, gender, and weight. Once the clusters have been calculated, validated through cross validation, and verified with appropriate metrics, the population statistics of each cluster will be identified and may demonstrate a particular group receiving different treatment. There may be various groups all being treated appropriately but in different manners, resulting in different clusters. SHAP, a technique that compares the relative contribution of a given feature to the output, has been used in similar instances to provide some context about each cluster and instance within (Louhichi et al. 2023). If there is a cluster that is found through post-analysis to have a large proportion of a minority group member within, then it may be evidence of bias. This would usually then require a blind medical board to determine if the treatment holds valid medical value. If the consensus is that it is not an appropriate (or a poorer quality) treatment plan, then the pattern could be used as evidence of bias. Each instance could also be traced back with proper authorization to make amends and give these patients a more appropriate treatment plan.

Many clustering models require a parameter indicating the number of expected clusters. To use such a model, researchers often prioritize hyperparameter tuning and provide various parameter values via a grid-search to uncover the model with the best verification metric value (Louhichi et al. 2023). However, there are clustering models that do not require this parameter, and these may be better suited as there is not a known or expected number of clusters sought (Fahim 2020). Both above modelling strategies will be implemented independently, the exact implementation remains undecided. Datasets may have many patterns hidden within them; it is essential that the right ‘channel’ is being observed through feature selection. To include features that have a poor information load would only introduce noise and result in weaker verification values and have potential result in overfitting (Dash & Liu 2000). To that end the following table indicates features believed to capture the pattern well (*Table 1*).

Table 1: Features containing information relevant to patterns of bias that may be useful to include in the model building phase.

Feature	Explanation
Admission type	How a patient appears to a hospital potentially indicates a socioeconomic component (Wild et al. 2010).
Discharge Disposition	How a patient left the hospital provides information about the care they were able to receive and/or some socioeconomic factors (Spooner et al. 2017).
Time in hospital	Indicative of the cost a hospital was willing to incur for a given patient as well as the patient's ability to afford that care (American Diabetes Association 2018; Kapoor et al. 2011; Niohuru 2023).
Medical speciality	The speciality of the attending physician can be grouped into 'relevant' and 'not relevant' to diabetes. This may relate to transfers, or be indicative of patients receiving incorrect or suboptimal care (Hashem, Chi & Friedman 2003; Singh & Venkataramani 2024).
Number of laboratory procedures	This indicates the cost of assets the clinician is prepared to incur or the ability of patients to afford that care. It may indicate complex cases, or the degree of attention, focus, and commitment received from the medical team.
Number of non-laboratory procedures	May capture instances when a clinician did not provide a patient with the proper care and/or may be indicative of neglect or a poorly equipped hospital for this condition.
Number of medications received	An indication of a patient's ability to afford medication and/or of the complexity of the case. This may also capture situations where a patient is not being given the best possible care.
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, lack of attention (particularly in conjunction with the number of procedures, etc), and more generally of system failure (Fraser et al. 2010).
Payer code	Indicates how the patient paid for the care (e.g. insurance, Medicare/Medicaid, out-of-pocket, etc.) and has socioeconomic implications (Kapoor et al. 2011).
Diabetic medication	Feature specifying if any diabetes medication was prescribed. This has been included as research has found a discrepancy in racial groups and socioeconomic groups when it comes to receiving preventative care (Piette et al. 2010; Pu & Chewning 2013).
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 simply didn't work for that individual (Auerbach et al. 2016).
Readmission	This indicates if a patient returned to hospital after their initial visit. Perhaps due to negligent or suboptimal care received initially (Auerbach et al. 2016).

Note: It is understood that these features do not necessarily indicate poor care or neglect, they do however have the capacity to harbour patterns of bias which is why they have been selected and explained as they have.

Dimension reduction techniques such as principal component analysis (PCA) can obtain components of the data that present the highest variance and assist in removing noise (Louhichi et al. 2023). Though this does risk making the clusters themselves less interpretable as the features are lost while the information remains (Dash & Liu 2000; Maćkiewicz & Ratajczak 1993). This may be practically useful if there is too much noise and complexity in the dataset. The dataset has several other features including the use of specific medications which may hold useful information and strengthen the analysis with a PCA approach.

Another key consideration is the distance metric being used, the way in which instances are considered to relate will have an enormous impact on how they are clustered (Kumar, Chhabra & Kumar 2014). There are many methods for managing mixed data, however they often required additional models to be implemented to give the categorical features numerical values (Ahmad & Khan 2019; Noorbehbahani, Mousavi & Mirzaei 2015). Another alternative is to use a model that can tolerate mixed data by handling categorical and numerical features differently, such as the k-prototype algorithm (Ahmad & Khan 2019). One option that appeals because of its simplicity is Gower's similarity coefficient this metric can manage mixed datasets. It takes the average of scaled distances variable by variable and gives a value between 0 and 1 (D'Orazio 2021). Though there are complications to using an 'unweighted' Gower's distance (as it treats all data types equally), it handles missing data and doesn't appear to require one-hot encoding or dummy variables (D'Orazio 2021). As there is no reason to believe one feature should be weighted more than any other to uncover a relevant pattern, the unweighted choice gives all features equal opportunity to present that information. The choice of verification metric is also extremely important as it dictates the weight of any evidence found. There may be difficulty in this analysis as there is much variability (i.e. in how an individual behaves, time of the year, etc.). While it was acknowledged that this may be reduced with PCA, the degree cannot be known ahead of experimentation. It is difficult then to determine how compact clusters may become, however it is obvious that there needs to be an adequate amount of separation between clusters and a measure of their homogeneity in relation to age, ethnicity, gender, and weight in post-analysis in order to be informative. Silhouette scores are often used to verify cluster separation, they work best with spherical clusters (Rousseeuw 1987). This metric has been used to verify cluster separation in similar research when values close to 1 are achieved on a scale of -1 to 1 (Shahapure & Nicholas 2020). Comparing the outcome to what other researchers have arbitrary determined to indicate a verified result may not be very meaningful given the degree of variability and the context of the problem are quite different from other obtainable articles.

Final remarks

Some may argue that this technique is not useful as it does not provide concrete evidence. However, when framed in an ethical light, the value of this research is much clearer. With the goal of protecting human life, improving efficiency, finding mechanisms to reduce cost, and promoting discussion on a contentious but perhaps ignored topic, there is value in the work. With the only costs involved being cheaper than the current alternative techniques (which are also fraught with verification issues), one could argue there is a moral obligation to try. The suggested methodology seeks to comment on potential bias in a way that does not introduce bias to the research itself. It draws on patterns recorded in historical datasets that exist in a system or organisation's culture. It attempts to offer evidence of bias in a non-accusatory manner that can lead to discussion, and perhaps result in collaborative change.

References

Adams, A, Soumerai, S, Lomas, J & Ross-Degnan, D 1999, 'Evidence of self-report bias in assessing adherence to guidelines', *International Journal for Quality in Health Care*, vol. 11, no. 3, pp. 187-192.

Ahmad, A & Khan, SS 2019, 'Survey of State-of-the-Art Mixed Data Clustering Algorithms', *IEEE Access*, vol. 7, pp. 31883-31902.

American Diabetes Association, A 2018, 'Economic Costs of Diabetes in the U.S. in 2017', *Diabetes Care*, vol. 41, no. 5, pp. 917-928.

Arnold, D & Whaley, C 2020, 'Who pays for health care costs? The effects of health care prices on wages', *The Effects of Health Care Prices on Wages (July 21, 2020)*.

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'.

Bai, G, Yehia, F & Anderson, GF 2020, 'Charity care provision by US nonprofit hospitals', *JAMA internal medicine*, vol. 180, no. 4, pp. 606-607.

Béland, D, Rocco, P & Waddan, A 2023, *Obamacare wars: Federalism, state politics, and the Affordable Care Act*, University Press of Kansas.

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.

Brown, LD 2003, 'Comparing Health Systems in Four Countries: Lessons for the United States', *American Journal of Public Health*, vol. 93, no. 1, pp. 52-56.

Caplan, RL 1989, 'The commodification of American health care', *Social Science & Medicine*, vol. 28, no. 11, 1989/01/01/, pp. 1139-1148.

Chua, K-P 2006, *SINGLE PAYER 101*.

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

D'Orazio, M 2021, 'Distances with mixed type variables some modified Gower's coefficients', *arXiv preprint arXiv:2101.02481*.

Dash, M & Liu, H 2000, 'Feature Selection for Clustering', in T Terano, H Liu & ALP Chen (eds), *Knowledge Discovery and Data Mining. Current Issues and New Applications,* Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 110-121.

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

Emanuel, EJ 2008, 'The Problem with Single-Payer Plans', *The Hastings Center Report*, vol. 38, no. 1, pp. 38-41.

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.

Fahim, A 2020, 'Finding the number of clusters in data and better initial centers for K-means algorithm', *International Journal of Intelligent Systems and Applications*, vol. 13, no. 6, p. 1.

Feagin, J & Bennefield, Z 2014, 'Systemic racism and U.S. health care', *Social Science & Medicine*, vol. 103, 2014/02/01/, pp. 7-14.

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.

Gale, A 2019, 'It's the Prices, Stupid: Why the United States is So Different from Other Countries', *Mo Med*, vol. 116, no. 1, Jan-Feb, pp. 6-7.

Galvan, MJ & Payne, BK 2024, 'Implicit Bias as a Cognitive Manifestation of Systemic Racism', *Daedalus*, vol. 153, no. 1, pp. 106-122.

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.

Hashem, A, Chi, MTH & Friedman, CP 2003, 'Medical errors as a result of specialization', *Journal of Biomedical Informatics*, vol. 36, no. 1, 2003/02/01/, pp. 61-69.

Health, HTHCSoP 2021, 'Injustice Laid Bare: A Pandemic’s Revelation in American Healthcare', Video, in @HarvardPublicHealth (ed.).

Hoffman, ED, Jr., Klees, BS & Curtis, CA 2000, 'Overview of the Medicare and Medicaid Programs', *Health Care Financ Rev*, vol. 22, no. 1, Fall, pp. 175-193.

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.

Kannan, VD & Veazie, PJ 2014, 'Predictors of Avoiding Medical Care and Reasons for Avoidance Behavior', *Medical Care*, vol. 52, no. 4.

Kapoor, JR, Kapoor, R, Hellkamp, AS, Hernandez, AF, Heidenreich, PA & Fonarow, GC 2011, 'Payment source, quality of care, and outcomes in patients hospitalized with heart failure', *Journal of the American College of Cardiology*, vol. 58, no. 14, pp. 1465-1471.

Kumar, V, Chhabra, JK & Kumar, D 2014, 'Performance evaluation of distance metrics in the clustering algorithms', *INFOCOMP Journal of Computer Science*, vol. 13, no. 1, pp. 38-52.

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.

Maćkiewicz, A & Ratajczak, W 1993, 'Principal components analysis (PCA)', *Computers & Geosciences*, vol. 19, no. 3, 1993/03/01/, pp. 303-342.

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

Miller, WD, Peek, ME & Parker, WF 2020, 'Scarce Resource Allocation Scores Threaten to Exacerbate Racial Disparities in Health Care', *CHEST*, vol. 158, no. 4, pp. 1332-1334.

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.

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.

Niohuru, I 2023, 'Healthcare Affordability', in I Niohuru (ed.), *Healthcare and Disease Burden in Africa: The Impact of Socioeconomic Factors on Public Health*, Springer International Publishing, Cham, pp. 105-120.

Noorbehbahani, F, Mousavi, SR & Mirzaei, A 2015, 'An incremental mixed data clustering method using a new distance measure', *Soft Computing*, vol. 19, no. 3, 2015/03/01, pp. 731-743.

Piette, JD, Heisler, M, Harand, A & Juip, M 2010, 'Beliefs about prescription medications among patients with diabetes: variation across racial groups and influences on cost-related medication underuse', *Journal of health care for the poor and underserved*, vol. 21, no. 1, pp. 349-361.

Pu, J & Chewning, B 2013, 'Racial difference in diabetes preventive care', *Research in Social and Administrative Pharmacy*, vol. 9, no. 6, 2013/11/01/, pp. 790-796.

Rousseeuw, PJ 1987, 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics*, vol. 20, pp. 53-65.

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

Selden, TM 2020, 'Differences Between Public And Private Hospital Payment Rates Narrowed, 2012–16', *Health Affairs*, vol. 39, no. 1, pp. 94-99.

Shahapure, KR & Nicholas, C 2020, 'Cluster Quality Analysis Using Silhouette Score', in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA),* pp. 747-748.

Singh, M & Venkataramani, A 2024, *Rationing by Race*, National Bureau of Economic Research.

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.

Spooner, KK, Salemi, JL, Salihu, HM & Zoorob, RJ 2017, 'Discharge Against Medical Advice in the United States, 2002-2011', *Mayo Clinic Proceedings*, vol. 92, no. 4, 2017/04/01/, pp. 525-535.

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.

Wild, SH, McKnight, JA, McConnachie, A & Lindsay, RS 2010, 'Socioeconomic status and diabetes-related hospital admissions: a cross-sectional study of people with diagnosed diabetes', *Journal of Epidemiology & Community Health*, vol. 64, no. 11, pp. 1022-1024.

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