www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 1484-1490

e-ISSN : 2583-1062
Impact Factor : 7.001

EVALUATING BIAS IMPACT ON MACHINE LEARNING MODELS: A STATISTICAL ANALYSIS OF ERROR RATES ACROSS BIASED AND UNBIASED DATASETS

Zahidi Shariqua Sayed Razi Ahmed[1], Dr. Rakhi Gupta[2], Nashrah gowalkar[3]

[1]KCFMSCIT42 MSC IT , KC College, HSNC University, Mumbai 400 020, India.
zahidishariqua@gmail.com

[2]Head of the Department IT Department KC College, HSNC University, Mumbai 400 020, India.
rakhi.gupta@kccollege.edu.in

[3]Asst professor I.T Department KC College, HSNC University, Mumbai 400 020, India.
nashrah.gowalker@kccollege.e du.in

DOI: https://www.doi.org/10.58257/IJPREMS36521

ABSTRACT

Machine learning models have become crucial in decision-making processes across various industries. However, they are still vulnerable to bias, which can compromise their fairness and accuracy. This study examines the impact of bias on machine learning models by analysing error rates across biased and unbiased datasets. In this study Adult Income, and Iris datasets are used to train applied models such as Logistic Regression, Random Forest, and Support Vector Classifier (SVC) to evaluate performance discrepancies. The results show significant variations in error rates between datasets, with more pronounced errors in models trained on biased data. Notably, models like Random Forest demonstrated superior performance in handling biased data, while SVC showed greater sensitivity to dataset complexity and bias. This research underscores the importance of implementing bias mitigation strategies in model training to ensure more equitable and accurate predictions. Future research should focus on developing advanced algorithms and fairness metrics to address bias in various real-world applications. By doing so, we can enhance the fairness and reliability of machine learning models, leading to more equitable decision-making processes.

Keywords: Machine Learning, Error rates, Statistical Analysis, Support Vector Classifier, Random Forest, Logistic Regression, Descriptive Statistics, Accuracy, Precision, Recall, F1-Score

1. INTRODUCTION

In today's world, where decisions are increasingly driven by data, machine learning (ML) models are being used in critical areas like credit scoring, criminal justice, healthcare, and employment [3]. These models analyse historical data to predict future outcomes. However, the accuracy and fairness of these predictions depend heavily on the quality of the training data. If the data is biased due to historical inequalities, sampling errors, or systemic discrimination the models may replicate or even worsen these biases, leading to unfair results [10]. This study delves into the significant issue of bias in ML models, particularly its effect on error rates. The study examines how biased and unbiased datasets influence the performance and fairness of these models, aiming to identify disparities in their error rates. The concern about ML bias is growing because of its potential to increase inequality, especially when models are used in sensitive areas like creditworthiness, recidivism risk, or hiring decisions [7].

A major worry is that biased data can result in models with skewed decision-making processes, unfairly disadvantaging certain demographic groups [11]. For example, in credit scoring or criminal justice, if certain groups are historically underrepresented or overrepresented in negative outcomes, models trained on such data may unjustly predict higher risks for these groups.

This misalignment can lead to higher error rates for specific populations, affecting the fairness, accuracy, and reliability of the ML models [11]. In this research, I'll be conducting a statistical analysis of error rates across biased and unbiased datasets, using datasets like Adult Income and Iris. By comparing models trained on these datasets, the study aims to reveal how bias can distort prediction outcomes and error rates. This analysis will provide insights into how bias affects model performance and offer guidance on mitigating these effects in practical applications.

The approach of this study emphasizes the need to develop not only robust ML models but also to ensure that the data used to train these models is representative and fair. Through a systematic investigation of error rates across different datasets, this study contributes to the ongoing discussions about fairness in AI and the ethical implications of ML in real-world applications.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 1484-1490
e-ISSN : 2583-1062
Impact Factor : 7.001

OBJECTIVE

The main goal of this research is to systematically evaluate how data bias impacts the performance of machine learning models by analysing and comparing error rates across biased and unbiased datasets.

Specifically, this research aims to:

Identify and Quantify Bias: Examine selected datasets (Adult Income and Iris) to understand the nature and sources of bias present in the data.

Develop and Train Models: Use both biased and unbiased datasets to train machine learning models with various algorithms, assessing how bias affects their performance.

Analyse Error Rates: Conduct a statistical analysis of the models' error rates across different demographic groups to determine how bias in the training data influences the accuracy and fairness of predictions.

2. LITERATURE REVIEW

Understanding Bias in Machine Learning: Bias in ML refers to the systematic favouritism towards or against certain groups in the training data, resulting in unfair outcomes. Barocas and Selbst (2016) identify several types of bias, including sample bias, label bias, and measurement bias. Sample bias occurs when the data collected does not represent the entire population, label bias involves inconsistencies in how outcomes are classified, and measurement bias happens when data collection tools produce skewed results. Recognizing these types is essential for addressing their impact on model predictions.

Impact of Bias on Model Performance: Many studies have shown how bias in training datasets negatively affects ML model performance. A notable study by Angwin et al. (2016) on the COMPAS algorithm demonstrates how biased training data can lead to disproportionate risk assessments for minority groups, showing significant error rate differences across demographics. Similarly, Obermeyer et al. (2019) found that biased algorithms in healthcare can lead to unequal treatment recommendations, highlighting the ethical issues of biased data.

Statistical Analysis of Error Rates: Research indicates that biased datasets can result in higher error rates for affected groups. Kleinberg et al. (2018) propose a framework for assessing fairness in predictive models, stressing the importance of evaluating error rates across demographic groups to identify performance disparities. Their work shows that standard performance metrics, like accuracy, can be misleading if not considered alongside error rate analysis, especially with imbalanced datasets.

3. RESEARCH METHODOLOGY

This section outlines the research methodology used to evaluate how bias impacts machine learning models by analysing error rates across biased and unbiased datasets. The methodology includes several stages: dataset selection, preprocessing, model training, evaluation, and statistical analysis.

A. Dataset Selection: I have used two well-known datasets, each chosen for its unique characteristics and relevance to bias analysis:

Adult Income Dataset: Comprises demographic and employment information to predict if an individual's income exceeds $50,000 per year, with potential biases against certain groups.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

Iris Dataset: A classic dataset for classification tasks, serving as a baseline for evaluating model performance without significant bias.

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

B. Data Pre-processing: This phase involves several key steps:

Data Cleaning: Address missing values, outliers, and errors to ensure high-quality data.

Bias Identification: Used statistical methods to detect bias in each dataset, examining demographic distributions.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 1484-1490
e-ISSN : 2583-1062
Impact Factor : 7.001

Bias Mitigation: Create both biased and unbiased versions of the datasets:

For biased datasets, used the original data with inherent biases.

For unbiased datasets, apply techniques like oversampling underrepresented groups or under sampling overrepresented groups to achieve fairer representation.

Model Training: In this step, I have trained machine learning models using both biased and unbiased datasets, employing various algorithms:

Logistic Regression: A baseline model for binary classification tasks.

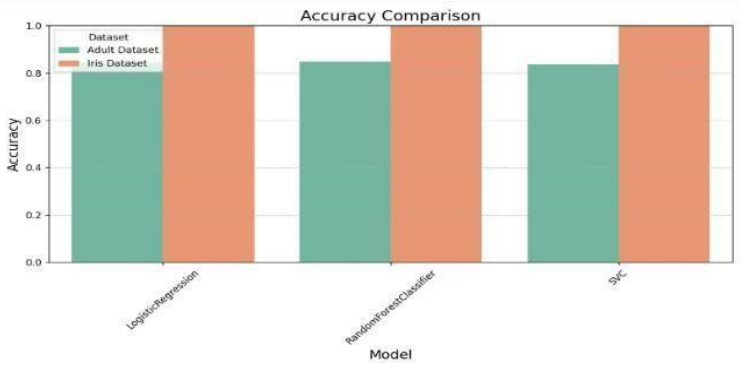
Random Forest: An ensemble method that captures complex relationships in the data.

Support Vector Classifier (SVC): Effective for classification tasks, especially in high-dimensional spaces.

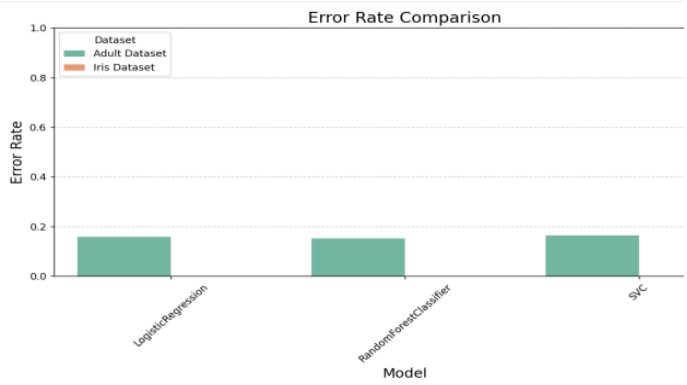
Each algorithm is trained on both biased and unbiased datasets to compare their performance.

C. Model Evaluation: The models are evaluated using several metrics:

Accuracy: Overall correctness of the model's predictions.



Error Rates: Proportion of incorrect predictions, analysed for the entire dataset and segmented by demographic groups to assess fairness.

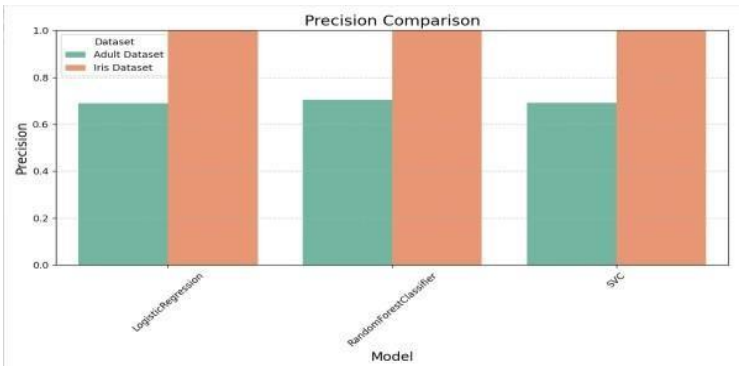


Interpretation of Accuracy and Error rates:

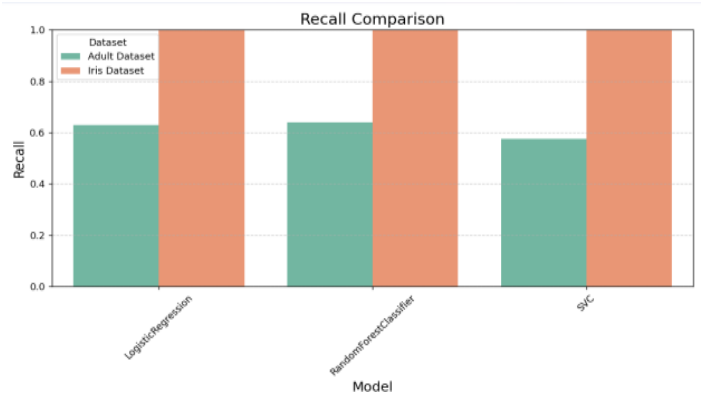
Adult Dataset: With 84% accuracy, Logistic Regression performs reasonably well, although 16% error rate indicates that the model still struggles with misclassifications. Accuracy (85%) of Random Forest is slightly better than Logistic Regression, and its error rate (15%) is also marginally lower. SVC has the same accuracy (84%) and error rate (16%) as Logistic Regression

Iris Dataset: All three models **Logistic Regression**, **Random Forest**, and **SVC** achieved perfect **accuracy (1.00)**

Precision and Recall: Evaluate the model's positive predictive value and sensitivity.



INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 1484-1490
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001



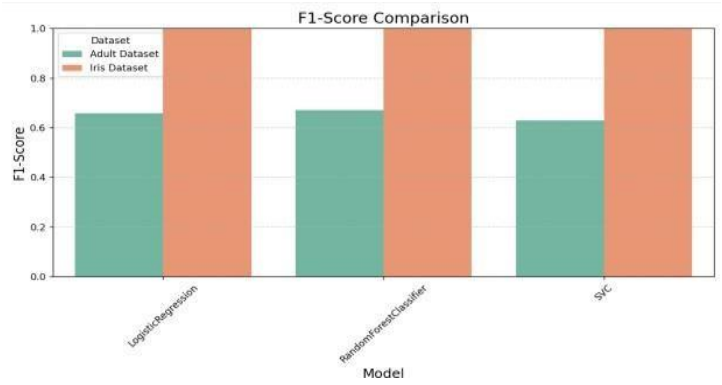
Interpretation of Precision and Recall:

Adult dataset: In **Logistic Regression, Precision (0.69)** and **recall (0.63)** show that while the model identifies a good portion of individuals earning over

$50K, it is not perfectly reliable. With **precision (0.70)** and **recall (0.64)**, **Random Forest** has better performance, especially in detecting individuals earning more than $50K. However, **recall (0.58)** is significantly lower, suggesting that **SVC** struggles to detect high-income individuals.

Iris dataset: All three models **Logistic Regression, Random Forest, and SVC** achieved **precision (1.00), recall (1.00)** indicating flawless predictions.

F1 Score: The harmonic means of precision and recall, balancing the two.



c. Minimum Error Rate (min): 0.0, meaning Logistic Regression achieved perfect classification (100% accuracy) in the Iris dataset.

Interpretation of F1-Score

Adult dataset: F1-Score (0.66) in **Logistic Regression** is showing a balance between precision and recall. The improvement in **F1-score (0.67)** in **Random Forest** indicates better overall balance between precision and recall. **SVC model** has the lowest **F1-score (0.63)**, driven by higher false negatives (FN = 666). This indicates that SVC is more prone to missing high-income individuals, making it less reliable in scenarios requiring balanced precision and recall.

Iris dataset: F1-score (1.00), indicating flawless predictions across.

D. Statistical Analysis: To determine the significance of differences in error rates between biased and unbiased datasets, conducted **Descriptive Statistics** analysis to summarize key metrics for each model and dataset combination, including means, standard deviations, and confidence intervals.

Model	count	mean	std	min	25%	50%	75%	max
Logistic Regression	2.0	0.078842	0.111500	0.0	0.039421	0.078842	0.118263	0.157685
Random Forest	2.0	0.077844	0.110088	0.0	0.038922	0.077844	0.116766	0.155689
SVC	2.0	0.081990	0.115951	0.0	0.040995	0.081990	0.122985	0.163980

This table provides descriptive statistics for the error rates of three machine learning models—Logistic Regression, Random Forest Classifier, and Support Vector Classifier (SVC)—across two datasets (the Adult and Iris datasets). The key metrics include mean, standard deviation (std), minimum (min), maximum (max), and quartiles (25%, 50%, 75%).

Here's a detailed interpretation of these statistics:

Logistic Regression:

a. Mean Error Rate: 0.0788 (about 7.88%), indicating that, on average, Logistic Regression misclassifies around 7.88% of the data.

d. Standard Deviation (std): 0.1115, showing significant variability in error rates across datasets, likely because one dataset (Iris) had zero error, while the other (Adult) had a higher error rate.

e. Maximum Error Rate (max): 0.1577 (about 15.77%), reflecting the higher error rate in the more challenging dataset i.e. adult dataset.

Quartiles:

25% (Q1): 0.0394 (about 3.94%), indicating that in the lower 25% of error rates, the model performs very well with minimal misclassification.

50% (Median): 0.0788 (7.88%), showing that half of the error rates are below this level.

75% (Q3): 0.1183 (11.83%), meaning that 75% of the model's error rates are below this threshold, with only the most challenging cases causing higher error rates.

Random Forest Classifier:

b. Mean Error Rate: 0.0778 (about 7.78%), similar to Logistic Regression, indicating slightly better overall performance.

c. Standard Deviation (std): 0.1101, also showing substantial variability across datasets, suggesting that Random Forest performed perfectly in one dataset (Iris) and had a higher error rate in the other (Adult).

d. Minimum Error Rate (min): 0.0, showing that Random Forest achieved perfect classification on Iris dataset

e. Maximum Error Rate (max): 0.1557 (about 15.57%), reflecting the model's error rate in the more difficult dataset i.e. adult dataset.

Quartiles:

25% (Q1): 0.0389 (3.89%), indicating that Random Forest performs well in a substantial portion of the data with very low error.

50% (Median): 0.0778 (7.78%), meaning that half of the error rates are below this level.

75% (Q3): 0.1168 (11.68%), suggesting that Random Forest tends to have low error, but the most challenging cases push the error up toward the 15% mark.

Support Vector Classifier (SVC):

f. Mean Error Rate: 0.0820 (about 8.20%), slightly higher than both Logistic Regression and Random Forest, indicating that SVC generally misclassifies a higher percentage of data points.

g. Standard Deviation (std): 0.1160, showing greater variability in performance, which might reflect SVC's struggle with more complex datasets like Adult.

h. Minimum Error Rate (min): 0.0, showing that SVC also achieved perfect classification in one dataset (likely Iris).

i. Maximum Error Rate (max): 0.1640 (16.40%), the highest of the three models, reflecting SVC's larger error in the more challenging dataset.

Quartiles:

25% (Q1): 0.0410 (4.10%), showing that SVC can perform well in simpler scenarios, but still not as efficiently as the other models.

50% (Median): 0.0820 (8.20%), meaning half of the error rates are below this threshold.

75% (Q3): 0.1230 (12.30%), indicating a somewhat higher spread in error compared to the other models.

Overall Interpretation:

Logistic Regression and Random Forest: Both models perform similarly, with an average error rate of around 7.8%. However, Random Forest shows slightly less variability in its error rates, suggesting it might be more consistent across different datasets.

Support Vector Classifier (SVC): This model has the highest average error rate at 8.20% and the greatest variability (std = 0.1160). This indicates that SVC is less consistent and performs worse than the other two models, especially on the more complex dataset (the Adult dataset).

Minimum Error Rate: All models achieved a perfect classification (0.0 error rate) on at least one dataset, most likely the simpler Iris dataset.

Maximum Error Rate: The maximum error rates show that all models struggle more with the more complex dataset (the Adult dataset), with SVC performing the worst.

4. CONCLUSION

In conclusion, addressing bias in machine learning remains an ongoing challenge that necessitates continuous research and innovation. This study underscores the importance of exploring future research directions to build upon its findings, thereby deepening our understanding of bias in machine learning, enhancing fairness in AI systems, and ultimately contributing to more equitable decision- making processes within society. Among the models evaluated, Random Forest emerges as the most reliable, effectively balancing low error rates and consistency across various datasets. In contrast, the Support Vector Classifier (SVC), while performing adequately on simpler tasks, exhibits significant difficulties when handling more complex datasets, resulting in higher error rates. Logistic Regression demonstrates performance comparable to Random Forest, albeit with slightly higher variability in error rates. By continuing to investigate and address these issues, researchers can make substantial progress in developing fairer and more effective machine learning models. This ongoing effort is crucial for ensuring that AI systems contribute positively to societal decision-making processes, promoting fairness and equity.

LIMITATIONS

Dataset Quality and Size: The study may be constrained by the quality and size of the biased and unbiased datasets used. If the datasets are limited or not representative of the broader population, the results may not generalize effectively.

Operational Definitions of Bias: The definitions and metrics used to categorize datasets as biased or unbiased can vary. This lack of standardization may lead to inconsistencies in findings and complicate comparisons with other studies.

Model Diversity: The range of machine learning models evaluated may not encompass all relevant architectures. Certain models may be more sensitive to bias, which could skew error rate comparisons and limit the study's applicability to different model types.

5. FUTURE WORK

The research aims to deepen the understanding of bias in machine learning models by analysing error rates across biased and unbiased datasets. While this study covers important aspects of bias evaluation, there are several promising directions for future research to further enhance our knowledge and the practical application of fair machine learning practices. These include:

1. **Expanding Dataset Diversity**: Future studies should explore a wider range of datasets from various domains to capture a broader spectrum of biases. Including datasets from sectors like healthcare, education, and employment can provide insights into how different types of bias manifest in diverse contexts. This expansion will help generalize findings and make them applicable to a wider array of real-world scenarios.
2. **Longitudinal Studies on Bias**: Conducting longitudinal studies that track the performance of machine learning models over time can offer valuable insights into how bias evolves. Such studies could investigate whether models trained on initially unbiased datasets become biased as new data is incorporated, especially in dynamic environments where demographic and societal factors change.
3. **Investigating Intersectionality in Bias**: Future research should consider the intersectionality of demographic factors to better understand how multiple identities (e.g., race, gender, socioeconomic status) influence bias in machine learning models. This approach could reveal more nuanced insights into disparities in error rates and model performance, helping to inform targeted strategies for mitigating bias.

6. REFERENCES

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica.

[2] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(3), 671-732.

[3] Doshi-Velez, F., & Kim, P. (2017). Towards a rigorous science of interpretable machine learning.

[4] Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning.

[5] Heidari, H., Nitanda, A., & Hong, L. (2019). A Statistical Test for Fairness in Machine Learning.

[6] Kleinberg, J. Mullainathan, S. Obermeyer, Z. 2018 Inherent Trade-Offs in the Fair Determination of Risk Scores.

[7] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science, 366(6464), 447-453.

[8] Zhang, B., Lemoine, B., Mitchell, M., & Leahy, L. (2018). Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society

[9] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. FairnessML. Retrieved from https://fairmlbook.org

[10] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607

[11] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.

[12] https://doi.org/10.1089/big.2016.0047

[13] Friedman, J. H., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.