An Investigation into Bias And Data Representativeness in Online Datasets

Hassan Najah
The Libyan Academy for Postgraduate Studies – Tripoli
21928505@Academy.edu.ly

Fathi Gasir
The Libyan Academy for Postgraduate Studies – Tripoli
Fathi.Elgasir @Academy.edu.ly

Keeley Crockett
Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK,
K.Crockett@mmu.ac.uk

Abstract— **The rapid growth of publicly available online datasets has created new opportunities for machine learning research; however, these datasets are often sampled from populations that are unknown to the researcher. As a result, the processes used for data collection, labelling, and sampling are frequently undocumented, increasing the risk of unrepresentative or biased samples. When training data fail to reflect the underlying population, model outputs may propagate or amplify existing biases. This paper investigates the presence of bias in online datasets derived from different domains. The research aims to identify biases associated with the distribution of attributes which represent personal protected characteristics and to evaluate methods that mitigate these biases by improving data representativeness. Three datasets were selected to be part of the study, each containing at least one protected variable and enabling binary classification. The protected variables examined included marital status, race, and gender, respectively. Model performance was assessed using accuracy, sensitivity, and specificity. To investigate and address bias, data quality and representativeness techniques were applied, as well as bivariate statistical analysis to remove variables with no significant association with the class label. Initial results showed that accuracy alone provided a misleading picture of model performance, particularly when sensitivity was low and specificity was high. The results indicated that after applying the data representativeness and mitigation techniques, the models achieved a more balanced performance across all three metrics, despite slight reductions in accuracy and specificity. The findings highlight that accuracy alone is insufficient as a performance metric. When datasets are not representative, bias mitigation methods that balance sensitivity and specificity may reduce accuracy but lead to more equitable outcomes. Classification models perform more reliably when class distributions are balanced, and fair systems should ensure equitable accuracy, sensitivity, and specificity across all protected subgroups.**

Keywords— Data Bias, Data Representativeness, Classification, Machine Learning, Fairness, Bias Detection, Dataset Imbalance, Dataset.

I. INTRODUCTION

Machine learning (ML) algorithms and statistical models are fundamental only as good as the data that is used to create them. Data bias and its impact on users is an important issue due to non-representative samples of the population [3]. Bias causes people discrimination based on sensitive or protected features, such as age, race, gender, religion and so on [4]. Decision-making in machine learning plays a big role in business nowadays, however; unlike people, machines do not become tired or bored. But, like people, algorithms can make biases in their decisions "unfair" [2]. Bias within ML and AI systems is where the data used to train, validate and test using different kinds of algorithms and the decisions come from. Bias can occur at many stages in the ML lifecycle. Algorithmic bias is often caused by how the data science team might label, collect, preprocess and sample and code the training data. Some specific causes include [5][6]:

- Biases in the training data: non-representative data is used to build the model. This includes the use of humans to label the dataset.
- Biases in algorithmic design: programing errors by the AI designer, it should better reflect the actual population regarding the sensitive (protective) attributes.
- Biases on evaluation: bad interpretation through how an individual or business applies the algorithm's output can lead to unfair outcomes depending on how they understand the outputs.

Some of the key challenges regarding data bias in online datasets may include:

1. Lack of Representation: Online datasets often overrepresent certain demographics while entirely excluding others. This results in sampling bias due to an unknown sampling method, where models make highly inaccurate predictions for underrepresented groups [7].

DOI: **10.5281/zenodo.20673267**

2. Data provenance: datasets compiled from historical online archives implicitly contain past human prejudices. Data provenance is primarily concerned with authenticity, providing details such as who created the data, the history of modifications and who made those changes. Furthermore, human annotators who label online data often introduce their own cognitive biases into the dataset [8].

3. Data Obsolescence: Information degrades over time, meaning online datasets may quickly become outdated and fail to reflect current realities [9].

4. Measurement Bias: Missing information and lack of documentation on how data was collected or labeled can lead to inconsistent metrics across different groups. If the measurement tools were geared toward specific cultural norms, the resulting data mischaracterizes real-world tasks [10].

The issues in this paper are constructed around three fundamental research questions:

RQ1: How can bias in protected variables, such as gender, ethnicity, and other demographic subgroups be systematically detected and addressed within machine learning datasets?

RQ2: What methods can effectively reduce biased data in web-based or online datasets to prevent unfair or discriminatory outcomes in automated decision-making systems?

RQ3: To what extent is model accuracy alone as an appropriate metric for evaluating machine learning models in the context of bias detection, and what additional metrics are required to ensure fair assessment?

To address these questions, a study was conducted which aims to identifying and mitigating bias in online open source datasets.

This research has three main contributions:

1. A step-by-step method for ensuring dataset representativeness that other researchers and practitioners can follow to reduce ethical bias in data-driven tasks.

2. A combined use of bivariate statistical analysis and machine-learning evaluation to remove irrelevant or misleading variables caused by poor dataset labelling.

3. An analytical insight into why balancing sensitivity and specificity is crucial, especially in relation to False Negative Rate (FNR) and False Positive Rate (FPR), and how a well-represented dataset maintains this balance without harming accuracy.

This paper is organised as follows: Section II presents related work, section III presents data analysis such as data description, transformation and datasets partitioning, section IV represents the study methodology, section V represents the results, VI represents the results discussion and finally section VII for the conclusion and future work.

II. Related Work

Machine learning bias refers to the tendency of an algorithm or model to produce systematically prejudiced or unfair outcomes due to erroneous assumptions in the learning process. It often arises from biased data, flawed model design, or external factors, and can lead to discrimination against individuals or groups. In computer science, this is also called algorithmic *bias* and is characterized by recurrent errors that result in “unfair” predictions or decisions [11].

In 2022, founding engineer Gabe Barcelos provided a detailed definition of machine-learning bias and its various forms [12]. He noted that high model accuracy does not necessarily indicate model stability or meaningful learning. Barcelos emphasized that bias could arise at any stage of the machine-learning lifecycle, including data collection, data preprocessing, feature engineering, data splitting or selection, model training, and model evaluation. In 2020, Crockett et., al. investigated gender differences in an automated deception detection system’s ability to classify non-verbal cues of deception when produced by males and females. The work utilized data from 36 head and facial nonverbal channels, which were represented in the form of image vectors. The patients were split into two separate sets, male group and female group. The classification of deceivers was done using machine learning methods on non-verbal cues. The results showed that gender did not affect the decision of classifiers, where both J48 default and Random Forest treated male and female subjects equally. The significance of these channels in decision making was also underlined in the studies reported in [3].

In addition, a survey by Mehrabi [2] examined ML biases and fairness in AI system design. The research highlighted that it was crucial to avoid biased treatment based on a particular group or populace. In their study, researchers examined real-world scenarios in which biases have been observed and highlighted that there is an urgent demand for better future research directions and solutions to combat the problem of biased AI models on wide-ranging benchmark Datasets.

A 2020 study by Ricardo Corrales-Barquero et al. [4] investigated gender bias mitigation in credit-scoring models. The authors conducted a systematic review across three major digital libraries—ACM Digital Library, IEEE Xplore, and SpringerLink—to identify the metrics and bias-mitigation techniques employed in credit-scoring research. Their analysis showed that most studies focused primarily on pre-processing approaches to reduce gender bias [4]. Andrejs S. defined how bias accidentally occurs when creating a dataset for any AI or machine learning project. highlighting the importance to data representation. He proposed an approach by establishing a set of criteria for assessing data representativeness [13]:

- Define the target population
- Choose an appropriate sampling method: random sampling, stratified sampling, cluster sampling, or convenience sampling.
- Check the data quality (pre-processing), identify and handle any missing values, outliers, duplicates, errors and imbalanced data.
- Evaluate data representativeness by using quantitative or qualitative methods, such as statistical tests, similarity metrics, or domain knowledge.

Although prior research [14] has demonstrated the presence of data bias in machine-learning outcomes and has defined bias in relation to legally protected attributes, there remains a substantial gap in formalising and measuring biases arising from unrepresentative data samples. Specifically, limited attention has been given to evaluating how data representation affects model fairness when the underlying dataset does not accurately reflect the target population. A critical gap persists in the development of robust evaluation frameworks and mitigation strategies capable of addressing intersectional representation bias. The research presented in this paper addresses this gap by focusing on sampling biases happen on account of selecting online datasets with unknown sampling methods of a population, causing some (sub) populations to be less likely to be represented, and what strategies to mitigate these biases plus, the paper addressed sampling method, lack of representation, resampling methods, data quality methods and feature selection methods regarding only data bias detection. While these not be new methods there is no published work on online datasets. The argument is that often researchers take these datasets and do not consider bias in relation to special characteristics.

III. Data Analysis

A. Data Description

Table I describes three online datasets available for researchers and training: Employee attrition [15], Adult income [16] and Health [17], with their variables. The three datasets meet the criteria for data classification with deferent businesses along with including at least one protected variable, the ability to apply classification process based on a binary class variable (the business) and the focus on variables that may impact the target variable.

Dataset Name	***Dimensions N * number of records***	***variables***	***Protected variable***	***Binary Class variable***
Employee Attrition [15]	17*1490	Attrition, Age, Distance from Home, Education, Environment Satisfaction, Gender, Job Involvement, Job Satisfaction, Marital Status, Monthly Income, Performance Rating, Relation Ship Satisfaction, Total Working Years, Work Life Balance, Years At Company, Years In Current Role, Years With Current Manager	Marital Status (divorced, married and single)	Attrition (yes, no attributes)
Adult Income [16]	12*43812	Age, Workclass, Education, Educational_num, Marital_status, Occupation, Relationship, Race, Gender, Hours_per_week, Native_country, Income	Race(white, black, other)	Income (>50K, =<50K attributes)
Health Dataset [17]	12*1190	Age, Gender, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting Electrocardiogram Results, Maximum Heart Rate Achieved, Exercise Induced Angina, Oldpeak =ST, STslope, Heart Attack	Gender (male, female)	Heart Attack (yes, no attributes)

TABLE I. DATASETS DESCRIPTION

B. Data Transformation

For each variable type, categorical data must be converted into factor variables. When categories represent an inherent order—such as rating scales or evaluation levels—they should be encoded as ordered factors to reflect their increasing sequence. This conversion process forms part of data transformation. In the employee attrition dataset [15],

variables' attributes such as "Yes" and "No" for the target variable "Attrition", should convert to 1 and 0, and the protected variables' attributes "Marital Status" (Single, Married and Divorced) should convert as levels into 1, 2, 3 respectively. Plus, some variables' attributes are ordered levels (Low, Medium, High and Very High) like in 'performance rating' variable as an example which are evaluated from 1-4, and from 1-5 for the variable "Education" and so on.

In the Adult Income dataset [16], categorical attributes are encoded numerically, with the target variable income represented as "1" for ">50K" and "0" for "≤50K." Similarly, protected attributes such as race (e.g., White and Black) are encoded as binary levels. In addition to binary features, the dataset includes ordinal variables—most notably education—which follow an inherent hierarchical sequence In the Health dataset [17], categorical variables such as the target attribute Heart attack are encoded binarily, with "Yes" represented as 1 and "No" as 0. Protected attributes, including Gender, are similarly encoded, with female and male assigned values of 0 and 1, respectively. The dataset also includes several ordinal features; for example, symptom severity levels (Mild, Moderate, Severe, Very Severe) are mapped to values 1–4, and resting electrocardiogram outcomes (normal, mild abnormality, hypertrophy) are encoded as values 1–3, with similar ordinal mappings applied to other clinically graded attributes.

C. Datasets Partitioning

Representative data—often referred to as a representative sample—is a subset of data that accurately reflects the characteristics of the larger population from which it is drawn [18]. In research, the sample comprises the individuals selected to participate in a study, whereas the population refers to the broader group to whom the results are intended to generalise. A representative sample is one that closely mirrors the attributes of the population and provides an unbiased depiction of its composition [19], [20]. Stratified sampling, a form of probability sampling, supports representativeness by dividing the population into homogeneous subgroups, or strata, based on relevant characteristics. Sample members are then randomly selected from each stratum, enabling comparison across groups and improving the accuracy of population estimates [21]. In the process of classification, it is essential to safeguard certain personal protected characteristic features from discrimination. These features are called protected variables or sensitive variables. Typical examples of protected variables include gender, ethnicity, age, religion, disability status, country of origin and more. Protected variables consist of attributes such as male, female, single, married and so on that are subject to consider when creating samples with subgroups or sub datasets [22].

In this study, each of the three datasets was disaggregated into subgroups based on the attributes of the protected variables. For the Employee Attrition dataset, the protected variables marital status was divided into three categories—single, married, and divorced—in addition to the overall population. Stratified sampling was then applied to evaluate whether each subgroup had an equal opportunity for comparable model performance [21].

For the Adult Income dataset, the data was partitioned according to the protected variables race, creating two subgroup datasets: White and Black. Model performance was assessed for each subgroup and compared with the performance on the full dataset. If a subgroup demonstrated similar performance to the overall dataset, this indicated a lower likelihood of bias affecting that protected attribute.

For the Health dataset, the protected variables gender was used to create male and female subgroup datasets. Performance metrics for each subgroup were compared with those of the overall dataset (Heart Attack classification). If the subgroup performance closely matched the group-level performance, the dataset was considered to exhibit low bias with respect to gender. Otherwise, further analysis would be required to distinguish between data bias and model (algorithmic) bias; however, the latter falls outside the scope of this paper.

IV. Study Methodology

Given the aim of identifying and mitigating bias in online datasets, this study proposes a comprehensive empirical procedure for examining bias in machine-learning classification systems. The investigation outlines a systematic approach for detecting specific sources of bias and applying best-practice mitigation techniques. Because the suitability of publicly available datasets cannot be predetermined, an experimental research design was adopted. Three diverse online datasets were selected, each containing at least one protected variable and a binary class label to enable classification. These datasets—originally collected for educational, research, and training purposes—were sourced from different populations, using unknown sampling strategies and domain-specific labelling practices. Consequently, it was not feasible to define a single target population or impose a uniform sampling approach. Instead, the analysis focused on applying data quality and data-representativeness techniques, supplemented by bivariate statistical analysis, to identify variables that may contribute to bias. The three datasets met the criteria required for binary classification tasks. Commonly used classification algorithms include support vector machines,

tree-based models, k-nearest neighbours, artificial neural networks, and logistic regression. In this study, decision trees, random forests, and logistic regression were selected because they are widely used across different application domains and provide sufficient diversity to enable meaningful comparison [23].

Fig I. provides a summary of the experimental methodology.

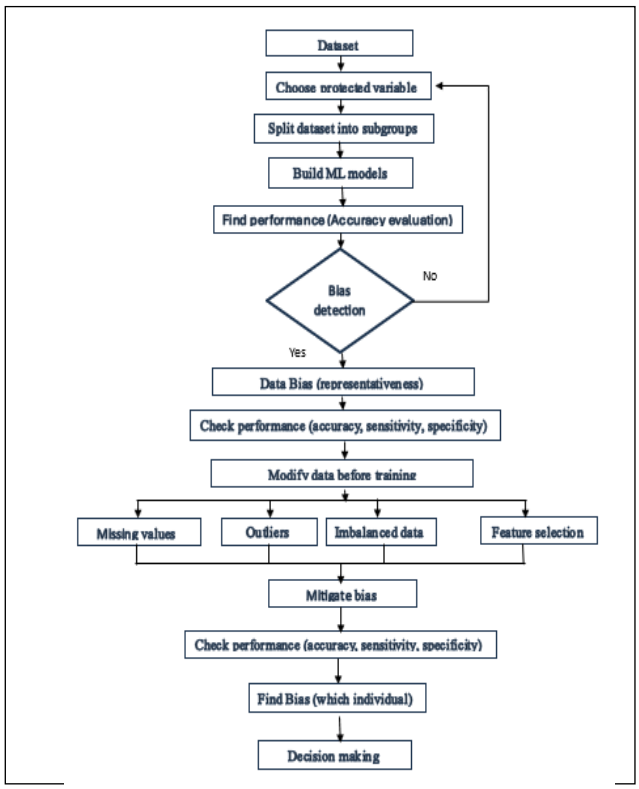

Fig 1. Methodology Structure

A. *Experimental Configuration and Parameter Settings*

1) *Experimental Configuration and Data Splitting:*

To ensure methodological consistency and reproducibility, all machine learning models in this study were implemented using the R programming environment with their default parameter configurations. Since the primary objective of this research is to investigate data representativeness and bias within the dataset rather than optimizing algorithmic structures, the default algorithm settings were maintained across all experiments.

Each dataset was randomly divided into training and testing subsets using a 70%–30% split. Randomization was applied during the sampling process to determine which observations were included in the training dataset and which were reserved for evaluation. Multiple experimental thresholds were used to explore different randomized splits of the data, allowing the identification of data representations that may produce more reliable predictive performance. Randomization plays a critical role in handling uncertainty, ensuring diverse solutions, and mitigating biases in algorithms [24].

2) *Model Implementation and Parameter Settings:*

Three machine learning models were implemented in this study: Decision Tree, Random Forest, and Logistic Regression.

The Decision Tree model was developed using the rpart package with the parameter method set to "class" to perform binary classification. The model employed the Gini impurity index as the default splitting criterion. Other parameters remained at their default values, including minsplit = 20 and minbucket = 7, maxdepth = 10 and Complexity Parameter (cp)= 0.01[25].

The Random Forest model was implemented using the randomForest package. The number of trees (ntree) was set to 100, which represents the default configuration for classification tasks. Other parameters, including the number of variables randomly sampled at each split (mtry = $\sqrt{p}$), with p being the number of predictor variables, were left at their default values[25][26].

Logistic regression was implemented using the (glm) function as a built-in function in R and can be used to fit generalized linear models with the binomial family and the (logit) link function, which represents the standard configuration for binary classification in R. All other arguments were retained at their default settings to ensure consistency across all models and datasets [25].

3) *Performance Evaluation :*

Model performance was evaluated using the confusion matrix function from the caret package to obtain multiple classification metrics. The parameter (mode) in R = "everything" was applied to compute a comprehensive set of performance measures beyond the default sensitivity and specificity. In addition, the parameter positive = "yes" was specified to explicitly define the positive class during model evaluation.

4) *Experimental Design :*

The experimental design methodology employed in this research was factorial design, which systematically tests combinations of hyperparameters of each model and investigates how multiple factors (independent variables) influence the target variable. Each factor is tested at distinct values, or levels, and the experiment includes every possible combination of these levels across all factors [27].

In the context of experimental design, a subject refers to an observation unit (e.g., data) to which treatments are applied, while a treatment represents a specific condition or set of factors applied to experimental units to measure its effect on a response variable [28]. In this research, the treatment

represents the impact of the factors (independent variables) on the target variable (the binary class) through the classification process, examining how this process achieves correct classification to avoid errors.

B. Initial Bias Detection

Since bias refers to the systematic error presents in a model's predictions that consistently skews results in a particular direction, one of the methods to detect bias in machine learning is to evaluate the performance of the model across different subgroups (strata) using the accuracy metric to identify any significant difference [3][29][30].

Accuracy = (True Positives + True Negatives) / (Total Predictions) Eq (1)

Accuracy alone is not a sufficient metric for detecting bias. As noted in prior work, high accuracy does not necessarily indicate that a model is stable or learning meaningful patterns, and accuracy can be particularly misleading when class distributions are imbalanced" [30][31]. The confusion matrix is therefore essential, as it highlights two critical types of classification errors—false negatives and false positives—which directly inform the assessment of model fairness and bias [32]. False Positive (FP) is where the model incorrectly predicted the positive class (Type I error).

False Negative (FN) is where the model incorrectly predicted the negative class (Type II error).

Accordingly, sensitivity and specificity are employed as complementary evaluation metrics to balance false-positive and false-negative errors, thereby mitigating the risk of type I and type II misclassifications. Using these measures enables the estimation of a more reliable, non-misleading accuracy, as they capture the distinct consequences associated with each form of error [33] [34] [35]:

Sensitivity (true positive rate) = True Positives / (True Positives + False Negatives) Eq (2)

Specificity (true negative rate) = True Negatives / (True Negatives + False Positives) Eq (3)

C. Mitigating Data Bias

Pre-processing techniques were applied as part of the data-representativeness criteria, including the handling of missing values, treatment of outliers, and implementation of resampling methods for datasets with imbalanced class distributions. In addition, feature-selection procedures, specifically bivariate analysis, were employed to eliminate variables that showed no statistically significant relationship with the target variable.

In statistical analysis, missing values arise when data are absent for some variables or cases, often due to limitations in the data-collection process. Handling missing data is essential because conventional statistical methods assume that every variable in a model is fully observed for all cases [36]. To address missing values in this study, several approaches were used. Rows containing missing text-based values were removed, while missing numeric values were imputed using the mean of the corresponding variable [37].

Table II shows how many records were removed due to missing values.

Dataset	**Number of rows**	**Number of rows with missing values**	**Number of rows without missing values**
Employee attrition [15]	1490	20	1470
Adult [16]	43812	12	43800
Health [17]	1190	0	1190

TABLE II. REMOVAL OF MISSING VALUES

Outliers are data points that lie outside the majority of the data in a particular data set. These values might be much higher or lower in value than other points and may impact the results of the data analysis in ways that misrepresent the data sample [38]. There are various techniques for identifying outliers, and in this study, the box plot method was utilized. The idea of box plot was presented by John Tukey in 1970 [39]. It displays key summary statistics such as the median, quartiles, and potential outliers in a concise and visual manner [40]. Table III shows the rows per dataset after removing outliers.

Datasets	**Number of rows before removing outliers**	**After removing outliers**
Employee attrition [15]	1470	1200
Adult [16]	43800	20309
Health [17]	1190	944

TABLE III. REMOVAL OF OUTLIERS

Class distribution refers to how instances are allocated across the different classes within a dataset. It indicates the number of examples belonging to each class and is crucial for assessing the reliability of classification models. Uneven or imbalanced class distributions can negatively affect algorithmic performance and may contribute to biased outcomes, particularly with respect to protected variables [40]. Imbalanced data problem is characterized by the classification of datasets exhibiting unequal class distributions, in other words, a big difference in the binary class distribution.

Table IV shows the class distribution for each dataset focusing on which target variable's categories are imbalanced in the three datasets.

Dataset	**Total records**	**Target variable (the class)**	**Class distribution (number of records for attributes)**	**Imbalanced class data?**
Employee attrition [15]	1200	Attrition	Yes = 219 No = 981	Yes
Adult [16]	20309	Income	<=50 k =15840 >5= k =4469	Yes
Health [17]	944	Heart Attack	Yes = 429 No = 517	No

TABLE IV. IMBALANCED DATA

Imbalanced class distributions within a dataset can be addressed through the use of resampling techniques. These techniques generally fall into two main categories: oversampling and under sampling. Oversampling aims to balance class sizes by increasing the number of instances in the minority class, either by duplicating existing samples or generating synthetic ones. This results in both classes having an equal number of samples, matching the larger class size. In contrast, under sampling reduces the size of the majority class to match the minority class. Although under sampling may risk discarding potentially useful information, it avoids introducing artificial or synthetic data and reduces the likelihood of creating inaccurate decision boundaries. In this case, both classes are balanced by adopting the smaller class size [41].

A central challenge in machine learning is selecting the most relevant features for building effective classification models. Feature selection aims to remove irrelevant or redundant variables, thereby improving model performance [42]. Numerous techniques exist, including statistical tests, recursive feature elimination, embedded methods, and feature-importance measures such as those derived from random forests [43]. In this study, the focus was on statistical testing through bivariate analysis, which evaluates the association between individual features and the target variable to identify variables that meaningfully contribute to model construction [44].

This study examines the relationship between the target variable and all other features, removing any variable that shows no significant association with the target. The choice of test depends on factors such as data type (categorical or numerical), sample size, and whether the analysis involves one or multiple samples. A p-value of less than 0.05 is used to indicate a statistically significant relationship with the target variable [45]. The appropriate statistical tests for this study are 1) *Wilcoxon test* used when the independent variable is numeric and the outcome is categorical; 2) Chi-square test used when the independent variable is categorical and the outcome is categorical [45].

A normality test is required for numerical variables to determine the appropriate statistical test for analysis. The Shapiro–Wilk test is commonly used to assess normality. If the test result is statistically significant (p-value ≤ 0.05), the variable is considered not normally distributed. A non-significant result indicates that the data do not significantly deviate from a normal distribution [46].

Table V shows the normality tests for all numeric variables in the three datasets.

Dataset	Numeric variable	Normality
Employee attrition [15]	Age, DistanceFromHome, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager	Significant
Adult [16]	Age, educational_num	Significant
	hours_per_week (Redundant values, all the data are same value)	
Health [17]	Age, Restingbps, Cholesterol, Maxheartrate, Oldpeak	Significant

TABLE V. NORMALITY TESTS

T-test applies on numeric variables when they are normally distributed, otherwise Wilcoxon test applies. Since all numeric variables in the three datasets were not normally distributed, Wilcoxon test has been used. On the other hand, Chi-square test applies on categorical variables as follows [44].

Table VI shows the results of the bivariate analysis for the Employee Attrition dataset [15].

Variable	Data type	Test	p-value	Impact on target variable 'attrition'?
Age	Num	Wilcoxon test	5.566e-08	Yes
DistanceFromHome	Num	Wilcoxon test	0.002312	Yes
MonthlyIncome	Num	Wilcoxon test	2.816e-09	Yes
TotalWorkingYears	Num	Wilcoxon test	1.559e-09	Yes
YearsAtCompany	Num	Wilcoxon test	7.494e-12	Yes
YearsInCurrentRole	Num	Wilcoxon test	4.174e-11	Yes
YearsWithCurrManager	Num	Wilcoxon test	1.259e-10	Yes
Education	Chr	Chi-square	0.5984	No
EnvironmentSatisfaction	Chr	Chi-square	6.847e-05	Yes
Gender	Chr	Chi-square	0.5394	No
JobInvolvement	Chr	Chi-square	1.174e-05	Yes
JobSatisfaction	Chr	Chi-square	0.0001657	Yes
MaritalStatus	Chr	Chi-square	6.505e-10	Yes
PerformanceRating	Chr	Chi-square	0.6249	No
RelationshipSatisfaction	Chr	Chi-square	0.06226	No
WorkLifeBalance	Chr	Chi-square	0.0005609	Yes

TABLE VI. BIVARIATE ANALYSIS FOR EMPLOYEE ATTRITION DATASET

Based on the p-values, which were greater than 0.05, four variables showed no statistically significant relationship with the 'Attrition' target variable and were therefore excluded from the dataset prior to model construction [43]. These variables were Education, Gender, PerformanceRating, and RelationshipSatisfaction which had no impact on the target variable 'attrition' unlike the other variables which will be involved in the model building due to their impact on the target variable 'attrition'. Following this feature-selection step, the three machine-learning

models (decision tree, random forest and logistic regression) were built after completing the full preprocessing pipeline, which included handling missing values, treating outliers, applying resampling techniques, and removing the non-significant variables. This enabled a more accurate assessment of bias and its mitigation through comparison of the resulting evaluation metrics.

Table VII shows the results of the bivariate analysis for the Adult dataset [16].

Variable	Data type	Test	P-value	Impact on target variable 'income'?
Age	Num	Wilcoxon	< 2.2e-16	Yes
Educational_num	Num	Wilcoxon	< 2.2e-16	Yes
Workclass	Chr	Chi-square	< 2.2e-16	Yes
Education	Chr	Chi-square	< 2.2e-16	Yes
Marital_status	Chr	Chi-square	< 2.2e-16	Yes
Occupation	Chr	Chi-square	< 2.2e-16	Yes
Relationship	Chr	Chi-square	< 2.2e-16	Yes
Race	Chr	Chi-square	< 2.2e-16	Yes
Gender	Chr	Chi-square	< 2.2e-16	Yes
Native_country	Chr	Chi- square	< 2.2e-16	Yes

TABLE VII. BIVARIATE ANALYSIS FOR ADULT DATASET

All variables had impact on the variable 'income' due to p-values less than 0.05 and will be involved in the models' building.

Table VIII shows the results of bivariate analysis – Health dataset [17]

Variable	Data type	Test	P-value	Impact on target variable 'Heart Attack'?
Age	Num	Wilcoxon	< 2.2e-16	Yes
Restingbps	Num	Wilcoxon	3.071e-06	Yes
Cholesterol	Num	Wilcoxon	0.0002275	Yes
Maxheartrate	Num	Wilcoxon	< 2.2e-16	Yes
Oldpeak	Num	Wilcoxon	< 2.2e-16	Yes
Gender	Chr	Chi-square	< 2.2e-16	Yes
Chestpaintype	Chr	Chi-square	< 2.2e-16	Yes
Fastingbloodsugar	Chr	Chi-square	0.0004327	Yes
Restingecg	Chr	Chi-square	1.756e-05	Yes
exerciseangina	Chr	Chi-square	< 2.2e-16	Yes
STslope	Chr	Chi- square	< 2.2e-16	Yes

TABLE VIII. BIVARIATE ANALYSIS FOR HEALTH DATASET

Since all variables have p-values less than 0.05, each shows a statistically significant association with the 'Heart Attack' outcome. Therefore, none of the variables will be removed from the dataset when developing the three machine-learning models.

V. RESULTS

This section shows the results of the bias detection and mitigation process using three models (decision tree, random forest and logistic regression) applied on the three datasets. Detection process applied as a pre-representativeness analysis (applied without removing missing values or dealing with outliers or applying bivariate analysis or even applying resampling methods). Mitigation process is the method applied on the three datasets dealing with missing values, outliers, bivariate analysis and resampling methods (over sampling and under sampling) as shown in the tables IX, X and XI and defined in the related work reference [13].

A. Employee Attrition Dataset

Table IX presents the results before applying data representativeness techniques and after dealing with missing values, outliers, removing non-significant variables and resampling the datasets (oversampling and under sampling).

Before applying any data-representativeness techniques, overall accuracy values appeared high across all three models; however, subgroup-level performance revealed disparities. For example, the Decision Tree achieved 85% accuracy for the full dataset but only 71% for single individuals, indicating potential bias. Similar patterns were observed with Random Forest (74% for single individuals versus 85% for the group) and Logistic Regression (70% versus 85%). These discrepancies suggest that accuracy alone masks meaningful differences in predictive performance across protected groups.

Sensitivity and specificity analysis further highlighted these biases. Sensitivity—the rate of correctly predicting attrition—was low across all subgroups before applying representativeness techniques, with values of 14% for the group, 27% for single individuals, 22% for married individuals, and 25% for divorced individuals. In contrast, specificity was very high, indicating that models frequently classified true attrition cases as non-attrition (false negatives). This imbalance, present across Decision Tree, Random Forest, and Logistic Regression models, demonstrates that a high overall accuracy can coexist with biased treatment of certain groups. A fair evaluation must therefore consider accuracy, sensitivity, and specificity together.

Algorithm	Decision tree				Random forest				Logistic regression			
Datasets	Group	Single	Married	Divorced	Group	Single	Married	Divorced	Group	Single	Married	Divorced

Pre-Representativeness	Accuracy %	85	71	87	84	85	74	88	87	85	70	86	86
	Sensitivity %	14	27	15	22	14	25	10	11	15	27	22	22
	Specificity %	97	91	95	91	97	96	97	95	97	89	93	93
Oversampling	Accuracy %	68	66	72	83	79	74	87	88	69	61	74	74
	Sensitivity %	59	47	23	42	20	29	14	14	76	41	57	57
	Specificity %	69	73	80	87	92	91	98	95	68	70	76	76
Undersampling	Accuracy %	59	61	53	62	64	66	64	69	67	58	49	49
	Sensitivity %	70	70	52	71	68	52	42	57	76	47	28	28
	Specificity %	57	67	53	65	63	71	67	70	65	63	51	51

TABLE IX. SUMMARY RESULTS BEFORE AND AFTER THE MITIGATION PROCESS FOR THE EMPLOYEE ATTRITION DATASET [11]

Applying oversampling to the imbalanced datasets, after addressing missing values, outliers, and irrelevant features, produced modest improvements in sensitivity for the Decision Tree and Logistic Regression models, but minimal change in Random Forest. These gains in true-positive detection came at the cost of reduced accuracy and specificity, illustrating the typical trade-off between balancing class detection and maintaining overall predictive performance.

Under sampling produced the most balanced outcomes between sensitivity and specificity, despite reducing accuracy. For the Decision Tree, subgroup performance became more consistent, particularly for single individuals compared with the group, suggesting improved fairness relative to the pre-processing stage. While Random Forest and Logistic Regression still showed some variation between groups, their subgroup performance improved compared with earlier results.

Overall, before applying representativeness techniques, substantial bias was expected, especially for single individuals, who exhibited both lower accuracy and very low sensitivity across all algorithms. Married and divorced individuals also exhibited bias driven by low sensitivity. After applying the full representativeness pipeline—especially under sampling—the models demonstrated improved fairness, with under sampling delivering the most balanced results across accuracy, sensitivity, and specificity, but the accuracy decreased (61% for single individual using decision tree, 53% for married and 62% for divorced) compared to 59% to the over all group indicating bias potential might occur on married individual not the single regardless of the low values of accuracy for all datasets. This change in the experiment indicates that the dataset Employee attrition was not well represented before the mitigation process. The mitigation process demonstrated that how it cannot be relied on the accuracy alone as a bias detection metric associated with data representativeness concepts. The same pattern of results is observed when using random forest and logistic regression models, as well as when applying oversampling methods.

B. Adult Dataset

Similarly, the results of the Adult dataset in Table X can be interpreted and discussed as follows:

Table X shows that overall accuracy was high across all models; however, subgroup comparison revealed disparities. In particular, the accuracy for Black individuals differed from both the group-level and White subgroup results, indicating potential bias. This shows that accuracy alone masked unequal model performance across demographic groups.

Algorithm		**Decision tree**			**Random forest**			**Logistic regression**		
Datasets		Group	White	Black	Group	White	Black	Group	White	Black
Pre-Representativeness	Accuracy%	82	81	88	83	82	88	82	82	88
	Sensitivity%	53	55	49	60	59	44	55	56	41
	Specificity%	92	90	95	90	90	95	92	91	97
Oversampling	Accuracy%	78	71	79	78	76	86	78	76	77
	Sensitivity%	80	83	83	79	80	70	85	82	78
	Specificity%	78	68	78	78	75	88	76	74	80
Undersampling	Accuracy%	77	76	76	76	76	76	78	76	77
	Sensitivity%	79	78	87	84	82	87	85	82	82
	Specificity%	77	75	75	74	74	75	76	75	76

TABLE X. SUMMARY RESULTS BEFORE AND AFTER DATA REPRESENTATIVENESS FOR ADULT DATASET[12].

After applying representativeness techniques, oversampling produced mixed results. Using Decision Tree and Random Forest, the accuracy values suggested possible bias toward White individuals. In contrast, Logistic Regression showed near-parity across the group and both subgroups, with accuracy values of 78% for the full dataset, 76% for White individuals, and 77% for Black individuals. Under sampling achieved an even stronger parity across all three models, suggesting a much lower likelihood of bias affecting either the White or Black subgroups.

Sensitivity and specificity analyses further clarified model behaviour. Before applying representativeness techniques, sensitivity values were substantially lower than accuracy

and specificity, indicating that the models failed to correctly identify positive cases across all groups. After applying resampling methods, a strong balance emerged between accuracy, sensitivity, and specificity, and accuracy did not decline significantly for any algorithm. All three models showed improved fairness; however, Logistic Regression performed particularly well under sampling, demonstrating balanced metrics across the group and subgroups—accuracy at 78%, 76%, and 77%; sensitivity at 85%, 82%, and 82%; and specificity at 76%, 75%, and 76%. These results indicate effective mitigation of bias following the application of representativeness techniques.

C. *Health dataset*

This dataset does not have class imbalance, so no resample method will be applied on the data. Table XI shows the three algorithms' results before and after dealing with only missing values, outliers and bivariate analysis.

Algorithm		Decision tree			Random forest			Logistic regression		
Datasets		Group	Male	Female	Group	Male	Female	Group	Male	Female
Pre-Representativeness	Accuracy %	82	84	78	93	94	91	86	86	85
	Sensitivity %	79	87	65	95	98	72	88	91	83
	Specificity %	86	79	95	90	85	97	84	77	86
After-Representativeness	Accuracy %	83	87	87	92	95	92	83	83	87
	Sensitivity %	84	89	70	89	95	70	80	82	80
	Specificity %	82	85	93	94	96	90	87	83	91

TABLE XI. SUMMARY RESULTS BEFORE AND AFTER DATA REPRESENTATIVENESS FOR HEALTH DATASET[13]

Random Forest and Logistic Regression produced similar accuracy values for the overall group and for the male and female subgroups, indicating a low likelihood of bias. In contrast, the Decision Tree model showed a notable disparity: accuracy was 78% for females compared with 82% for the group and 84% for males. This suggests that the model may exhibit bias affecting male individuals. After applying representativeness methods, all three algorithms showed improved consistency, and accuracy values became closely aligned across the group and subgroups.

Sensitivity and specificity analyses further clarified these patterns. Prior to applying representativeness techniques, sensitivity and specificity were generally balanced across models, with the exception of the Decision Tree, which showed a low sensitivity rate of 65%. Following the application of representativeness techniques, subgroup balance—particularly for females—improved. Logistic Regression demonstrated the strongest performance for bias detection and mitigation, achieving the most consistent balance across accuracy, sensitivity, and specificity.

D. *Results' evaluation*

A two-sample t-test was conducted to compare the classification accuracy of the subgroups in the three datasets to the overall groups. The tests considered the results of the three algorithms (decision tree, random forest and logistic regression) in a single test for each dataset as the purpose of these tests was to evaluate bias detection and mitigation outcomes resulting from the experiments.

In employee Attrition dataset performance comparison applied between (single, married and divorced) accuracy results and the Group accuracy as shown in table XII.

	Group vs Single		Group vs Married		Group vs Divorced	
Method	**P-value**	**Sig**	**P-value**	**sig**	**P-value**	**sig**
Pre-Representativeness	0.00039	Yes	0.066	No	0.492	No
Oversampling	0.389	No	0.393	No	0.048	Yes
Undersampling	0.643	No	0.199	No	0.035	Yes

TABLE XII .T-TEST RESULT FOR EMPLOYEE ATTRITION DATASET.

The results revealed a significant difference between "single" performance and the "group" performance ($p=0.00039<0.05$) before applying data representativeness. Data mitigation using both oversampling and undersampling method revealed a statistically significant difference between "divorced" performance and "group" performance ($p=0.048$ and $p=0.035$ respectively) this time indicating the correctness in the algorithm's predictions and how using the accuracy as a performance metric without considering the balance between the sensitivity and the specificity in a non-represented dataset. The chance of bias is still high and shifted from "single" to "divorced" and might occur on each individual due to a very low accuracy.

In Adult dataset the comparison results were between subgroups (white and black) and the overall group as shown in table XIII:

	Group vs White		Group vs Black	
Method	**P-value**	**sig**	**P-value**	**sig**
Pre-Representativeness	0.230	No	0.00007	Yes
Oversampling	0.093	No	0.384	No
Undersampling	0.158	No	0.382	No

TABLE XIII .T-TEST RESULT FOR ADULT DATASET.

The results revealed a significant difference between "black" performance and the "group" performance ($p=0.00007$) before applying data representativeness. Data

mitigation revealed no statistically significant difference between the subgroups performance and "group" performance indicating improvement in the performance and mitigating the chance of bias on both individuals.

In Health dataset the comparison results were between subgroups (white and black) and the overall group as shown in table XIV:

	Group vs male		Group vs female	
Method	**P-value**	**sig**	**P-value**	**sig**
Pre-Representativeness	0.833	No	0.660	No
After-Representativeness	0.669	No	0.503	No

TABLE XIV .T-TEST RESULT FOR HEALTH DATASET

T-test results revealed no significant differences before and after the mitigation processes indicating a very well represented dataset

VI. DISCUSSION

The results of this research stress the point "that high performance is not always the indicator that a model is stable and learning appropriately" [31]. A model can have high accuracy for the dominant class or group while performing very poorly for some minority subgroups, e.g., high accuracy and very low sensitivity (high false negative) in 'marital status' observed in the Employee attrition dataset, Table IX After applying resampling methods, data quality and bivariate analysis, accuracy and specificity decreased, and sensitivity increased resulting in a balance between sensitivity and specificity which made a sense in the evaluation process.

Potential bias was observed across all marital-status subgroups in the Employee attrition dataset, in contrast to the initial accuracy-only evaluation, which suggested bias solely for single individuals. For example, Table IX shows that under sampling with the Decision Tree, the lowest accuracy shifted from the single subgroup to the married subgroup, indicating a different pattern of possible bias. The Random Forest results highlighted divorced individuals as potentially affected, while Logistic Regression continued to suggest bias among single individuals. These variations demonstrate that relying on accuracy alone produces an incomplete and sometimes misleading assessment of subgroup bias.

In the Adults dataset Table X, accuracy and specificity were high across all algorithms, but sensitivity remained comparatively low. For instance, the Decision Tree achieved 82% accuracy and 92% specificity but only 53% sensitivity. The Black subgroup showed consistently lower accuracy than both the full group and the White subgroup, indicating a potential bias in all three algorithms. After applying mitigation techniques, all models showed a slight decrease in accuracy but achieved a more balanced relationship between sensitivity and specificity. Logistic Regression, particularly under the under sampling method, produced the most balanced and stable results across demographic groups.

For the Health dataset Table XI, resampling was not required because the classes ("Heart Attack: Yes/No") were not imbalanced. All algorithms achieved high accuracy, and both Logistic Regression and Random Forest showed minimal differences between the group, male, and female subgroups. The Decision Tree displayed a small performance difference for the female subgroup. After applying representativeness techniques, accuracy values changed slightly but remained high across all models. Importantly, a good balance between sensitivity and specificity was observed both before and after mitigation. All three algorithms performed well overall, with Logistic Regression showing the most consistent and equitable results.

T- sample tests applied to evaluate the experiment results. The tests confirmed the difference between a non-represented dataset and a well-represented dataset by. Employee attrition dataset was non-represented, and the mitigation techniques revealed actual performance, but it was still non-represented so far as there are many reasons behind it, how data collected, labeled, sampled and more. Mitigating Adult dataset improved the performance and decreased the chance of bias to the extent saying it was well-represented. Health dataset was very well-represented where the mitigation did not change the bias level and the performance evaluation between the individuals and the overall group.

VII. CONCLUSION AND FUTURE WORK

This study presents a comprehensive empirical analysis for identifying and mitigating bias in machine-learning classification systems. The research addresses key questions concerning the detection of latent and systematic biases arising from under-represented online datasets, particularly where such biases manifest across protected demographic attributes and risk discriminatory outcomes in automated decision-making. It further examines which mitigation strategies are most appropriate at different stages of the machine-learning pipeline and questions the sufficiency of model accuracy as a standalone performance metric for bias detection and evaluation.

Methodologically, the work adopts a principled bias analysis grounded in subgroup-level investigation and data representativeness. Three widely used classification algorithms—Decision Tree, Random Forest, and Logistic Regression—are evaluated across three distinct online datasets. Bias is systematically assessed by measuring performance disparities between demographic subgroups,

enabling comparative analysis across both models and data contexts.

The findings demonstrate that bias in classification systems is predominantly data-driven. Experimental results indicate that the most effective interventions occur at the source of bias, revealed through analyses of data representativeness. Addressing data quality issues—including missing values, outliers, class imbalance, and skewed distributions via bivariate analysis—was shown to have a greater impact on reducing discriminatory performance gaps than model-level adjustments alone. These results underscore the limitations of relying solely on accuracy metrics and highlight the necessity of integrating data-centric bias diagnostics into machine-learning evaluation frameworks.

The importance of multi-metric evaluation: "Accuracy" is not enough and can be misleading. Metrics like sensitivity and specificity give a more well-rounded view of model performance on different subgroups. A fair system should balance such metrics for all protected subgroups, plus when the dataset is not well represented, the mitigation process that makes sensitivity and specificity in balance sense decreases the accuracy of the model. Classification models performed more reliably when class distributions are balanced. No matter what algorithm was used as most common classification algorithms are subject to the decision maker choice but after applying many experiments.

The evidence from this investigation suggested that:

- The Employee attrition dataset after the mitigation process showed a bad data representation occurs when information is stored or visualized in a way that is misleading and inconsistent resulting in low accuracy, sensitivity and specificity values even after bias mitigation process.
- The Adult dataset after mitigation process showed a not bad data representation due to a good accuracy, sensitivity and specificity results.
- The Health dataset after mitigation process showed a very good and stable data representation, high accuracy, sensitivity and specificity results plus the reason of the target variable ‘heart attack’ was balanced.
- Achieving "zero bias" in machine learning is generally considered impossible in real-world applications, even with well-represented or high-quality datasets. While a dataset can be balanced (equal representation of classes), it is rarely truly unbiased, as bias is defined by systematic errors or deviations from objective reality and some protected demographic characteristics might be biased in whatever machine learning or AI web-based project.
- Statistical tests evaluated bias detection and bias mitigation results and confirmed the interpretation of the results regarding how data representativeness standards can improve module’s performance and decrease the chance of bias.

The study framework that presented in this paper is a concrete step-by-step process that data scientists and engineers can follow to thoroughly audit their web-based machine learning projects for bias however it did not cover all aspects of data bias detection and mitigation.

Future research will aim to investigate automated bias detection and mitigation pipelines web-based applications and extended examination to additional datasets and protected attributes, plus try different classification algorithms.

REFERENCES

[1] Ghiffari Assamar Qandi, Nur Aini Rakhmawati,2020,” A survey of Fairness in Machine Learning for Indonesian General Election Research”, 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), IEEE, https://ieeexplore.ieee.org/document/9274583.

[2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan, 2019, “A Survey on Bias and Fairness in Machine Learning ”, Cornell University, https://arxiv.org/abs/1908.09635.

[3] K. Crockett, J. O’Shea and W. Khan, "Automated Deception Detection of Males and Females from Non-Verbal Facial Micro-Gestures", 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207684.

[4] Ricardo Corrales-Barquero, Gabriela Marín-Raventós, Elena Gabriela Barrantes, “A Review of Gender Bias Mitigation in Credit Scoring Models”, 2021 Ethics and Explainability for Responsible Data Science (EE-RDS), IEEE, 2021, https://ieeexplore.ieee.org/document/9708589.

[5] Markus Spiske, “Algorithmic bias”, Greenlining website, https://greenlining.org/wp-content/uploads/2021/04/Greenlining-Institute-Algorithmic-Bias-Explained-Report-Feb-2021.pdf.

[6] Alexandra Jonker, Julie Rogers, “What is algorithmic bias?”, IBM, 20/9/2024, https://www.ibm.com/think/topics/algorithmic-bias.

[7] Alexandra Olteanu, Carlos Castillo , Fernando Diaz , Emre Kıcıman, “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries”, National Library of Medicine, PMID: 33693336, 2019, https://pmc.ncbi.nlm.nih.gov/articles/PMC7931947/

[8] Tim Mucci, “What is data provenance?”, IBM, https://www.ibm.com/think/topics/data-provenance.

[9] Patrick O Halloran, “Common Data Quality Challenges and How to Overcome Them“, WhereSpace, 2024, https://www.wherescape.com/blog/the-importance-of-maintaining-data-quality/.

[10] Amandalynne Paullada and others, “Data and its (dis)contents: A survey of dataset development and use in machine learning research”, ScienceDirect, Volume 2, Issue 11, 12 November 2021, 100336, https://www.sciencedirect.com/science/article/pii/S2666389921001847.

[11] Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A. and Kyriazis, D., 2024. Bias in machine learning: A literature review. Applied Sciences, 14(19), p.8860, https://doi.org/10.3390/app14198860.

[12] Gabe Barcelos, “Understanding Bias in Machine Learning Models”, Arize AI platform, 2022, [online], Available: https://arize.com/blog/understanding-bias-in-ml-models.

[13] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” Journal of Big Data, vol. 6, no. 1, pp. 1–54, 2019, doi: 10.1186/s40537-019-0192-5.
[14] Nima Shahbazi, and others,”Representation Bias in Data: A Survey on Identification and Resolution Techniques”, ACM digital library, July 2023, https://dl.acm.org/doi/10.1145/3588433#:~:text=Representation%20Bias%20in%20data%20can,work%20within%20their%20respective%20domains.
[15] Kagle, https://www.kaggle.com/code/faressayah/ibm-hr-analytics-employee-attrition-performance.
[16] Kagle, https://www.kaggle.com/code/nikszodape/adult-income-dataset-with-knn.
[17] https://archive.ics.uci.edu/datasets.
[18] Market research solutions, “How to get a representative sample for your survey”, https://www.surveymonkey.com/market-research/resources/how-to-get-a-representative-sample/.
[19] Y. Tillé and A. Matei, “Basics of Sampling for Survey Research,” in The SAGE Handbook of Survey Methodology, C. Wolf, D. Joye, T. W. Smith, and Y.-C. Fu, Eds. London, U.K.: Sage Publications, 2016, pp. 309–334..
[20] Intellectus-Consulting. “What is a Representative Sample?”, https://www.statisticssolutions.com/what-is-a-representative-sample/
[21] VOXCO, “Stratified Sampling vs Cluster Sampling”, https://www.voxco.com/blog/stratified-sampling-vs-cluster-sampling/.
[22] Zhen Peng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman, 2024,"Fairness Improvement with Multiple Protected Attributes: How Far Are We?", In 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE ’24). https://arxiv.org/pdf/2308.01923.
[23] Data camp: https://www.datacamp.com/blog/top-machine-learning-use-cases-and-algorithms?utm_cid=22660585401&utm_aid=181540422635&utm_campaign=230119_1-ps-other~dsa-tofu~blog_2-b2c_3-nam_4-prc_5-na_6-na_7-le_8-pdsh-go_9-nb-e_10-na_11-na&utm_loc=9000887-&utm_mtd=c&utm_kw=&utm_source=google&utm_medium=paid_search&utm_content=ps-other~nam-en~dsa~tofu~blog~machine-learning&gad_source=1&gad_campaignid=22660585401&gbraid=0AAAADQ9WsETu1LWiSI0UQxVbhhorNE4v&gclid=Cj0KCQiA-YvMBhDtARIsAHZuUzLsl18djTvvaL_ucvxVziRslPE2lyrX2-3l77K65haqJbH5oRcD-gkaAsDVEALw_wcB.
[24] B. R. P. M. Basnayake and N. V. Chandrasekara, "Application of Random Data Splitting Technique for Time Series Modeling Using Feedforward Neural Network and Support Vector Regression Models," Journal of Applied Data Sciences, vol. 7, no. 2, pp. 1496-1515, May 2026, doi: 10.47738/jads.v7i2.1216..
[25] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Evaluating Fairness Strategies in Educational Data Mining: A Comparative Study of Bias Mitigation Techniques ," *Electronics*, vol. 14, no. 9, p. 1856, May 2025. doi: 10.3390/electronics14091856.
[26] P. Probst, M. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *arXiv preprint arXiv:1804.03515*, 2018.
[27] G. A. Lujan-Moreno, G. A. Lujan-Moreno, P. Howard, O. Rojas, and D. C. Montgomery, “Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study,” Expert Systems With Applications, vol. 109, pp. 195–205, Nov. 2018, doi: 10.1016/J.ESWA.2018.05.024.
[28] Iván Carrascosa, “Choosing the Right Experimental Design: A Decision Tree Approach”, STATOLOGY, July 2025, https://www.statology.org/choosing-the-right-experimental-design-a-decision-tree-approach/
[29] Cody Blakeney, Gentry Atkinson, Nathaniel Huish, Yan Yan, Vangelis, Metsis, Ziliang Zong, “Measuring Bias and Fairness in Multiclass Classification, 2022 IEEE International Conference on Networking, Architecture and Storage (NAS), IEEE, 2022, https://ieeexplore.ieee.org/document/9925287.
[30] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, G. O. R. Cruz, R. M. Peixoto, G. A. S. Guimarães, L. L. Santos, M. M. Araujo, M. Cruz, E. L. S. Oliveira, I. Winkler, and E. G. S. Nascimento, "Bias and Unfairness in Machine Learning Models: A Systematic Literature Review," *arXiv preprint*, 2022.. https://arxiv.org/pdf/2202.08176
[31] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Evaluating Fairness Strategies in Educational Data Mining: A Comparative Study of Bias Mitigation Techniques," Electronics, vol. 14, art. no. 1856, May 2025, doi: 10.3390/electronics14091856.
[32] Z. Lin, Z. Hao, and X. Yang, “Effects of Several Evaluation Metrics on Imbalanced Data Learning”, doi: 10.3969/j.issn.1000-565x.2010.04.027.
[33] Aniruddha Bhandari, “Confusion Matrix in Machine Learning”, https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/.
[34] R. Trevethan, “Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice,” *Frontiers in Public Health*, vol. 5, p. 307, Nov. 2017, doi: 10.3389/fpubh.2017.00307.
[35] Jacob Shreffler; Martin R. Huecker., “Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios”, National library of Medicine, https://www.ncbi.nlm.nih.gov/books/NBK557491/.
[36] Paul D. Allison," Missing Data ", Statistical Horizons, statistical training website, https://statisticalhorizons.com/wp-content/uploads/2012/01/Milsap-Allison.pdf
[37] L. Ren, T. Wang, A. Sekhari, H. Zhang, and A. Bouras, “A Review on Missing Values for Main Challenges and Methods,” Information Systems, vol. 119, Art. no. 102268, Oct. 2023, doi: 10.1016/j.is.2023.10226
[38] Coursera Staff, “What Are Outliers in Data Sciences?”, https://www.coursera.org/articles/what-are-outliers.
[39] Etibar Aliyev, “How can you measure data quality for ML? “, AI and the LinkedIn community, https://www.linkedin.com/advice/3/how-can-you-measure-data-quality-ml-skills-machine-learning-bsjbc#:~:text=Measuring%20data%20quality%20for%20machine,aligns%20and%20doesn't%20contradict.
[40] Riccardo Bellazi, Blaz Zupan, “Class Distribution”, Science Direct, https://www.sciencedirect.com/topics/computer-science/class-distribution.
[41] Annie Kim, Inkyung Jung, "Optimal selection of resampling methods for imbalanced data with high complexity",2023, PLoS ONE 18(7): e0288540. https:// doi.org/10.1371/journal.pone.0288540.
[42] S.Visalakshi , V. Radha ,“Literature Review of Feature Selection”, 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, https://ieeexplore.ieee.org/document/7238499.
[43] Okan Bulut, “Effective Feature Selection: Recursive Feature Elimination Using R”, Towards Data Science, 2025, https://towardsdatascience.com/effective-feature-selection-recursive-feature-elimination-using-r-148ff998e4f7.
[44] Neha pokhriyal, Shashi Kant Verma, “Statistical Feature Extraction/Selection for Small Infrared Target”, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India, https://ieeexplore.ieee.org/document/7732355
[45] Aman, “ Feature Selection in Machine Learning”, Analytics Vidhya, 01 May, 2025, https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/.
[46] S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for Normality (Complete Samples),” Biometrika, vol. 52, no. 3/4, pp. 591–611, Dec. 1965, doi: 10.2307/2333709.