1. **Introduction:**

Whereas artificial intelligence has brought a lot of difference in various industry sectors, it has also brought to the fore the issue of sampling bias that may undermine the fairness and accuracy of the whole AI system. If the data used for training is not representative enough, the resultant models will be biased and inequitable. As such, the problem is of actual importance in developing fair and reliable AI. The paper dwells on strategies for finding and mitigating sampling bias to bring forth more accurate AI predictions with fairness.

1. **Literature Survey:**

Sampling bias arises when a sample is not representative of the population it was taken from, leading to biased interpretation. It can take many forms such as selection bias, where the sample is chosen in a way that is not random at all, self-selection bias wherein individuals choose themselves to be part of a group and under coverage bias which involves failing to properly represent some members of the population in the sample.

The identification techniques for sampling bias consist of comparing the features of samples with known parameters in populations using statistical exams like chi-squared test which examines how well observed values fit expected ones. Some ways of addressing this problem are through random sampling where every element in the population has an equal chance of being included, stratified sampling whereby respondents are divided into subgroups and allotted proportionally, and post-stratification where weights given to samples reflect characteristics of population after data collection. Sampling biases affect research by undermining findings because of systematic errors hence limiting their application to wider populations (Panzeri et al., 2008).

The negative implications of bias in AI include the amplification of existing inequalities, whereby such effects fall more heavily on the marginalized groups, and the general denial of access to services that sustain daily life, in particular, health and finance prestige. The AI systems can stereotype gender further and present the differences between areas, for example, in security systems and job opportunities. More broadly, however, biased AI shapes societal norms and cements discriminatory narratives further impairing efforts toward equality. Biased AI not only means race, gender, age, or able-bodied/differently-abled discrimination; it also puts the onus on the developer, company, and government for ensuring fairness and transparency. It brings in distrust amongst the general public towards adapting new technologies and closes the amount of individual freedom further, going on to enhance contemporary power dynamics.

AI bias mitigation strategies include steps for making the training data representative in preprocessing through oversampling, under sampling, and synthetic data generation; care while choosing models for analysing the data; adjustments in post-processing to handle biases and ensure fairness; audit trails capturing dataset bias and augmentation processes; adversarial debiasing, which trains models to be unbiased to certain types of bias. The challenges of mitigating bias are fivefold: diverse and representative training data, the identification and measuring of different types of bias, potential trade-offs between fairness and the accuracy of AI systems, ethical considerations of priority in the kinds of bias that have to be attended to, and ethical considerations on priority in groups for mitigation efforts (Ferrara, 2023b).

The Fair-SMOTE technique is an oversampling method designed to balance datasets by generating new data points that match the distribution of parent points. This process uses K-nearest neighbours to create new data points, ensuring similarity in Boolean, symbolic, and numeric columns. The goal is to repeat this process until subgroup sizes are balanced, thereby achieving a fairer training dataset. Following Fair-SMOTE, Fair Situation Testing identifies and removes biased data points, allowing the model to be trained on the remaining fairer data without significantly affecting performance.

Performance metrics for Fair-SMOTE are evaluated against default, optimized pre-processing (OP), and Fairway methods. Fair-SMOTE demonstrates better performance in reducing bias and improving accuracy, recall, and F1 score across datasets such as Adult Census, Compas, and German Credit. The advantages of Fair-SMOTE include its effectiveness in balancing datasets and reducing bias for multiple protected attributes (Chakraborty et al., 2021).

Bias in datasets often arises due to unrepresentative sampling, leading to skewed model training. Existing mitigation techniques, such as data augmentation, re-weighting, and fairness constraints, can be complex and computationally expensive. A proposed method to address this issue is the Bias Mimicking Approach. This concept involves mimicking the bias present in the original data distribution to create a more balanced dataset. The technique employs a simple sampling strategy to adjust the dataset distribution, ensuring better representation of minority groups.

In the experimental setup, various publicly available datasets with known bias issues were used, and evaluation metrics included fairness and performance measures such as accuracy, F1 score, and bias reduction measures. Results and analysis showed that the bias mimicking approach significantly improved fairness metrics while maintaining or improving overall model performance. When compared to existing bias mitigation techniques, this approach was found to outperform them in terms of simplicity and effectiveness (Qraitem et al., 2022).

**3. Methodology:**

Four different sample bias mitigation strategies—**SMOTE**-, **INPROCESSING**-, **RANDOMSAMPLING**-, and **HYBRID** approaches—will be put into practice and assessed. We'll apply each strategy to our models and evaluate them using measures like accuracy and demographic parity difference to compare how well they performed. The purpose of this comparison analysis is to ascertain how well each technique reduces sampling bias and enhances model fairness and generalization.

**Identifying Sampling Bias**

**Descriptive Statistics**:

To comprehend the distribution of demographic categories within the dataset, we conducted a preliminary study. This entails figuring out what percentage of each race and gender there are in the dataset.   
For instance, we looked at the distribution of people by sex (Male, Female) and race (White, Black, Asian-Pacific-Islander, Other).

**Class Imbalance**:

The target variable (income\_>50K) showed an imbalance, with most cases falling into one class (income <=50K). Predictions may become skewed as a result of this imbalance since the model may come to favour the majority class.

**Fairness Metrics**:

Using fairness criteria like Equalized Odds Difference and Demographic Parity Difference, we assessed the initial model's fairness. These measurements aid in determining whether model outcomes differ significantly between various sensitive groups.

**In-Processing Method**: Techniques applied during the training of the machine learning model to mitigate bias. These methods modify the learning algorithm itself to ensure fairness constraints are met.

**Hybrid Method**: A blend of several methods to deal with bias at various phases of the pipeline for machine learning. Here, it encompasses both pre-processing (changing the dataset prior to training) and in-processing (changing the model or training procedure) methods.

**Grid Search Method**: A methodical approach to hyperparameter tuning that iteratively explores various combinations of parameter values, cross-validating each one to ascertain which combination yields optimal performance.

**Demographic Parity Constraints**: A type of fairness constraint used to make sure the model's predictions are not influenced by sensitive characteristics like gender or race. Every demographic group should have comparable rates of favourable outcomes.

**Demographic Parity Difference**: A fairness statistic that measures the difference in positive outcome rates for groups that are characterized by sensitive attributes. Reduced values signify a lower degree of bias.

**Adversarial debiasing**: A technique that aims to identify and reduce biases by training an adversary model alongside the primary prediction model. When the adversary successfully predicts sensitive traits (such race or sexual orientation) from the main model's predictions, the main model suffers a penalty, which incentivizes the main model to become less biased.

**SMOTE (Synthetic Minority Over-sampling Technique):** SMOTE creates synthetic instances of the minority class in order to balance a dataset. Instead of just repeating the current data, it produces a more balanced dataset by interpolating between minority class examples that already exist and their closest neighbours.

**Random Sampling:** In random sampling, instances are either removed at random from the majority class to undersample it or oversampled the minority class by randomly replicating instances. Although this method aids in dataset balancing, it occasionally causes overfitting or the loss of significant data.

**4. Dataset Preparation:**

The [UCI Adult Income](https://archive.ics.uci.edu/dataset/2/adult) dataset contains a number of socioeconomic and demographic characteristics specific to Americans. The following characteristics are included: age; work class (private, self-employed, etc.); final weight (sampling weight); education (preschool to doctorate), education-num (years of education); marital status (married, single); occupation (tech support, craft repair, etc.); relationship status (husband, own child, etc.); race (White, Black, etc.); sex (male or female); capital gain, capital loss; hours per week worked; and native country (e.g., United States, Mexico).

The dataset is prepared by loading the data, handling missing values, and applying one-hot encoding to categorical variables. Features and target variables are defined, followed by splitting the data into training and testing sets and standardizing the features. Sensitive features, specifically race and sex, are extracted for analysis

**5. Results:**

**SMOTE (Synthetic Minority Over-sampling Technique):**

SMOTE was used to oversample the minority class in order to balance the dataset. The **accuracy obtained with this method was 61.27%.**

There was a fair distribution of predictions for both genders, as evidenced by the comparatively small **demographic parity difference for sex (5.40%).** The model still appeared to have a significant amount of racial bias, though, as the **demographic parity gap for race was much greater at 30.57%.**

**In-Processing Method:**

Using the Fairlearn library, a logistic regression classifier is employed with the Grid Search method to find the best model that satisfies demographic parity constraints, ensuring predictions are independent of sensitive features.

A number of predictors are assessed for accuracy and fairness, and the model that best balances the two is chosen.

With demographic parity differences of **17.01% for sex** and **23.22% for race**, the best model yielded an **accuracy of 84.72%**.

**Hybrid Method:**

A blend of multiple methods to address bias at various points in the pipeline of machine learning. Here, it encompasses both pre-processing (changing the dataset prior to training) and in-processing (changing the model or training procedure) methods.

Disparities in the distribution of these sensitive traits were found during the first data exploration process, and these differences were further measured using measures of demographic parity difference.

The test **accuracy of our model was 84.12%**. The results of the demographic parity difference for **race and sex were 20.8% and 17.7%**, respectively, demonstrating that prejudice persisted even after debiasing efforts.

**Random Sampling:**

Through the random under- and over-sampling of the majority and minority classes, random sampling was utilized to balance the dataset.

An **accuracy of 79.21%** was obtained with this method. Gender projections showed a slight bias, as evidenced by the **5.36%** discrepancy in demographic parity for sex. The **difference in** **demographic parity for race was 11.02%.**

**6. Conclusion:**

To address sample bias in our machine learning models, we tested four distinct bias mitigation strategies in this study: SMOTE, in-processing, hybrid approach, and random sampling. Accuracy and disparities in racial and gender demographic parity were used to evaluate each approach.

SMOTE: **Accuracy of 61.27%** was attained. Significant racial bias was shown by the relatively low **demographic parity difference for sex (5.40%)** and the considerable **demographic parity difference for race (30.57%).**

In-Processing Method: At **84.72%, this method produced the best accuracy**. However, there was considerable bias in both categories as evidenced by the **demographic parity differences of 17.01% for sex and 23.22% for race**.

Hybrid Method: **84.12% accuracy** was attained. **The racial and sexual demographic parity differences were 20.8% and 17.7%**, respectively, indicating that bias remained after applying several debiasing techniques.

Random Sampling: An **accuracy of 79.21%** was obtained using this method. It showed a fair reduction of bias across both sensitive traits, with the lowest **demographic parity difference for sex (5.36%)** and a moderate **demographic parity difference for race (11.02%).**

The results showed that while the in-processing method had the best accuracy (84.72%), it still showed some bias in terms of the disparities in demographic parity across sex and race. Although the accuracy of the hybrid approach was 84.12%, bias was not considerably decreased.  
  
**Random sampling** emerged as a balanced approach, showed the lowest demographic parity gap for sex (5.36%) and a moderate reduction in racial prejudice (11.02%), while obtaining a relatively high accuracy (79.21%).

When accuracy and fairness are taken into account, **Random Sampling** is the most effective technique. It offers a reasonable balance between fairness and model performance, hence minimizing bias in predictions for both gender and race.

**7. References:**

Panzeri, S., Magri, C., & Carraro, L. (2008). Sampling bias. *Scholarpedia Journal*, *3*(9), 4258. https://doi.org/10.4249/scholarpedia.4258

Ferrara, E. (2023b). Fairness and Bias in Artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Chakraborty, J., Majumder, S., & Menzies, T. (2021). *Bias in machine learning software: why? how? what to do?* https://doi.org/10.1145/3468264.3468537

Qraitem, M., Saenko, K., & Plummer, B. A. (2022, September 30). *Bias Mimicking: A Simple Sampling Approach for Bias Mitigation*. arXiv.org. https://arxiv.org/abs/2209.15605

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.