SigmaRed

Table of Contents – AiSCERT LITE – Brief user manual

SigmaRed

Introduction

The Aiscert Lite Platform, offered by SigmaRed, performs an accelerated and automated AI bias analysis. It provides insights regarding the extent of Bias in each AI algorithm through various Bias metrics and graphs.

This document provides a simple walkthrough of the Aiscert Lite platform including initiating a new project, uploading data, and conducting bias analysis through its functionalities.

Contact Details for onboarding support

Email - Contact@sigmared.ai

Alternative Email – Murali@sigmared.ai , Vijaya@sigmared.ai,

Phone - + 1 646 670 1262

Key features of AiSCERT Lite platform

The following are the key features of this platform. More details of each feature are given in the walk-through section below.

1. Uploading Data
2. Defining protected attributes
3. Performing Bias assessment
4. Generating Bias Report

SigmaRed

Description of Features

This section describes the features of the platform.

1) Creating a New Project:

- After logging into the platform, the “AI Bias Projects” page is displayed. The user can create a new project using the “Create New AI Bias Assessment” button.
- Click on the "Create New AI Bias Assessment" button to initiate a new project.
- Enter the project name (mandatory) and provide project details as required by the form.
- Click the "Next" button to proceed.

2) Uploading Data:

- In the "Upload Data" section, choose a data file (CSV format).
- Ensure that the CSV file contains at least one protected attribute.
- Click "Submit" to upload the data.

3) Project Setup and Defining Protected Attributes:

- On the next page, identify the protected attributes from the features of the dataset and mention them on the “Project Setup” form.
- The “Project Setup” form consists of five menus: Target, Protected Attribute, Privileged, Unprivileged, and Predicted column.
- The menu “Target” is a drop-down menu in which the actual Y column from the uploaded CSV file will appear. Select that from the Target drop-down menu.

- In the “Protected Attribute” drop-down menu, all the columns from the CSV file other than the actual Y and predicted Y column will appear as checkboxes. The user can select multiple options from the “Protected Attribute” drop-down menu.
- According to the options the user selected from the “Protected Attribute”, the user has to type the values in the Privileged and Unprivileged textboxes.
- For Example, If the user wants to add “Race” and “Sex” columns of CSV file as Protected Attributes, he/she has to select those checkboxes from the Protected Attribute drop-down menu which may contain other column names in the options like Region, Language etc.,
- The user needs to input the values of Privileged and unprivileged text boxes for which the bias metrics will be calculated based on the protected attributes selected by the user. For example, the user should enter "Male" and "Female" in the Privileged and Unprivileged text boxes, in any order they choose if they have selected “Sex” as protected attributes.
- The user can type as many values as the selected protected attributes in the privileged and Unprivileged text boxes separated by comma.
- For Example, Sex and Race are the selected protected attributes, the user has to give “Male”, “Asian” in the Privileged text box and “Female”, “American” in Unprivileged text boxes.
- The menu “Predicted Column” is a drop-down menu in which Predicted Y column from the uploaded csv file will appear. Select that from the Predicted drop-down menu.
- The terminology like Protected Attributes, Privileged values, Unprivileged values are explained in APPENDIX – Key Terminology

4) Responsible AI Analytics:

- Once the Project Setup is done after feeding the values, click on the "Submit" button.
- This will take you to the next page for analyzing bias.

5) Bias Summary:

- In the Bias Summary page, a pie chart displays Low, Medium, and High scores indicating bias metric impact overview.
- All bias metrics are listed in a table along with their results, description, and normal range.
    - S.No
    - Metric Name
    - Metric Description
    - Metric Value
    - Criticality
    - Normal Range for each metric
- Around 30 bias metrics are calculated and displayed here. A brief description of these metrics is provided in Appendix - A.

6) Bias Metrics:

- In the Bias Metrics page, all the bias metrics are displayed in bar charts.
- Only one value is shown in each graph, as this platform provides analytics on live data when the platform features are involved, and hence previous data is not displayed.

7) Feature-wise Evaluation:

The Feature-wise Evaluation page displays additional metrics and charts, as applicable providing further insights.

APPENDIX – Key Terminology

Bias: A systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage.

Fairness metric: A quantification of unwanted bias in training data or models.

Favourable label: A label whose value corresponds to an outcome that provides an advantage to the recipient. The opposite is an unfavourable label.

- Favourable Classes: - Positive Class
- Un-Favourable Class: - Negative Class

Protected attribute: An attribute that partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion. Protected attributes may vary from case to case as required by the customer's AI model.

Privileged protected attribute: A value of a protected attribute indicating a group that has historically been at a systematic advantage.

Privileged Groups: - Groups that are highly privileged or of high importance in protected features. Ex: Race / Ethnicity column, then ‘White’

Unprivileged Groups: Groups that are less important than privileged groups.
Ex: Race column, then ‘Black’

APPENDIX – A- Descriptions for Metrics

1. True Positive (TP)

True Positives = Number of data points that are positive and also predicted by the given AI model, as positive

True positives refer to the number of correctly identified positive cases in a given situation. In other words, it represents the instances where a test or a model correctly identifies something as positive when it is actually positive.

For example, if a medical test correctly identifies 90 out of 100 people with a certain disease as having the disease, those 90 cases would be considered true positives. It is an important measure to assess the accuracy and effectiveness of tests or models in correctly identifying positive outcomes.

2. False Positive (FP)

False Positives = Number of data points that are negative but predicted by the given AI model, as positive.

False positives refer to the number of incorrect positive identifications in a given situation. It means that a test or a model wrongly identifies something as positive when it is actually negative.

For example, if a medical test incorrectly identifies 10 out of 100 healthy people as having a certain disease, those 10 cases would be considered false positives.

3. True Negative (TN)

True Negatives = Number of data points that are negative and also predicted by the given AI model, as negative

True negatives refer to the number of correctly identified negative cases in a given situation. It represents the instances where a test or a model correctly identifies something as negative when it is actually negative.

For example, if a medical test correctly identifies 80 out of 100 healthy people as not having a certain disease, those 80 cases would be considered true negatives.

4. False Negatives (FN)

False Negatives = Number of data points that are positive but predicted by the given AI model, as negative

False negatives refer to the number of incorrect negative identifications in a given situation. It means that a test or a model wrongly identifies something as negative when it is actually positive.

For example, if a medical test incorrectly identifies 20 out of 100 people with a certain disease as not having the disease, those 20 cases would be considered false negatives.

5. True Positive Rate (TPR)

$$\mathrm{TPR} = \mathrm{TP}/\mathrm{P}$$

Out of all the positives how many were predicted as positive.

In simpler terms, TPR tells us how good a test or model is at catching positive cases. For example, if a medical test has a TPR of 90%, it means that out of 100 people with a certain disease, the test correctly identifies 90 of them as positive.

6. False Positive rate (FPR)

$$\mathrm{FPR} = \mathrm{FP}/\mathrm{N}$$

Out of all negative points how many points are falsely predicted as positive.

For example, let's say we have a medical test for a certain disease. If the FPR of the test is 5%, it means that out of 100 healthy individuals who take the test, 5 of them would receive a positive result even though they don't have the disease.

7. False Negative Rate (FNR)

$$\mathrm{FNR} = \mathrm{FN}/\mathrm{P}$$

In this formula, FNR represents the False Negative Rate, FN represents the number of False Negatives, and P represents the total number of Positive cases.

For example, let's say we have a medical test for a certain disease. If the false negative rate of the test is 10%, it means that out of every 100 people who actually have the disease, the test will incorrectly identify 10 of them as not having the disease.

7. True Negative Rate

$$\mathrm{TNR} = \mathrm{TN}/\mathrm{N}$$

To understand it better, let's consider a medical test for a certain disease. The true negative rate tells us how often the test correctly identifies healthy individuals as negative. For example, if the true negative rate is 90%, it means that out of 100 healthy people, the test correctly identifies 90 of them as negative.

8. Positive Predicted Values (Precision)

$$\mathrm{PPV} = \mathrm{TP}/\ \{\mathrm{TP}+\mathrm{FP}\}$$

Positive Predicted Value (PPV), also known as precision, refers to the proportion of correctly predicted positive cases out of all the predicted positive cases. In simpler terms, it measures the accuracy of a test or model in correctly identifying positive outcomes.

For example, let's say a medical test predicts that 100 people have a certain disease. Out of those 100 predicted positive cases, if 80 people actually have the disease, then the PPV or precision would be 80%. This means that 80% of the predicted positive cases are truly positive.

9. False Discovery Rate

$$\mathrm{FDR} = \mathrm{FP}/\{\mathrm{TP}+\mathrm{FP}\}$$

False Discovery Rate (FDR) is a statistical measure that helps us understand the proportion of false discoveries among all the discoveries made. In simpler terms, it tells us how many of the positive findings we have are actually false.

Imagine you are conducting a scientific study and you want to identify significant results. The FDR helps you determine the likelihood that some of the significant findings you identify are actually false positives.

For example, let's say you conduct a study and find 100 significant results. The FDR would tell you how many of those 100 significant results are likely to be false discoveries. If the FDR is 5%, it means that around 5 of those 100 significant results are expected to be false positives.

10. False Omission Rate

$$\mathrm{FOR} = \mathrm{FN} / \{\mathrm{TN}+\mathrm{FN}\}$$

False Omission Rate (FOR) is a measure that tells us how often a test or a model incorrectly fails to identify something as positive. In other words, it represents the proportion of cases where the test or model wrongly classifies something as negative when it is actually positive.

For example, if a medical test has a false omission rate of 10%, it means that out of 100 positive cases, the test fails to identify 10 of them as positive. This can lead to missed diagnoses or incorrect conclusions about the presence of a condition.

11. Negative Predicted Value (NPV)

$$\mathrm{NPV} = \mathrm{TN}/\{\mathrm{TN}+\mathrm{FN}\}$$

Negative Predicted Value (NPV) is a measure that tells us how reliable a negative test result or prediction is. It helps us understand the probability that a negative test result is actually accurate.

Let's say you take a medical test to check for a certain disease, and the test result comes back negative. The NPV tells you the likelihood that you truly don't have the disease, based on the negative test result.

For example, if a test has a high NPV of 95%, it means that out of 100 people who test negative, 95 of them are truly disease-free. On the other hand, if a test has a low NPV of 70%, it means that out of 100 people who test negative, only 70 of them are actually disease-free, while the remaining 30 may have the disease despite the negative result.

12. Accuracy

$$\mathrm{Accuracy} = \{\mathrm{TP}+\mathrm{TN}\} / \{\mathrm{P}+\mathrm{N}\}$$

Accuracy, in simple terms, refers to how correct or accurate something is. In the context of tests or models, accuracy measures how well they can correctly identify or predict outcomes. It is a way to assess how reliable and trustworthy a test or model is in providing the right results.

For example, let's say you have a medical test that is designed to detect a certain disease. If the test has an accuracy of 90%, it means that out of 100 people tested, it will correctly identify 90 people who have the disease and correctly identify 90 people who do not have the disease.

13. Error rate

$$\text{Error Rate} = \{FP + FN\} / \{P+N\}$$

The error rate is a measure that tells us how often a test or a model makes mistakes in identifying positive and negative cases. It is calculated by adding the number of false positives (incorrectly identified positives) and false negatives (incorrectly identified negatives), and then dividing that sum by the total number of positive and negative cases.

In simpler terms, the error rate gives us an idea of how often the test or model gets it wrong.

14. True Positive Rate Difference

True Positive Rate Difference = (TPR (unprivileged group) – TPR (privileged group))

This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.
Fairness for this metric is between - 0.1 and 0.1

15. False Positive Rate Difference

False Positive Rate Difference = (FPR (unprivileged group) – FPR (privileged group))
This metric is computed as the difference of false positive rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group.

Fairness for this metric is between -0.1 and 0.1

16. False Negative Rate Difference

False Negative Rate Difference = (FNR (unprivileged group) – FNR (privileged group))

This metric is computed as the difference of false negative rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group.

Fairness for this metric is between -0.1 and 0.1

17. False omission rate difference

False Omission Rate Difference = (FOR (unprivileged group) – FOR (privileged group))

This metric is computed as the difference of false omission rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group.
Fairness for this metric is between -0.1 and 0.1

18. False Discovery rate difference

False Discovery Rate Difference = (FDR (unprivileged group) – FDR (privileged group))

This metric is computed as the difference of false discovery rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group.

Fairness for this metric is between -0.1 and 0.1

19. False positive rate ratio

False Positive Rate Ratio = FPR Unprivileged / FPR Privileged

Ratio of the FPR metric of Privileged and unprivileged groups.

This metric is computed as the ratio of false positive rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

20. False Negative Rate Ratio

False Negative Rate Ratio = FNR Unprivileged / FNR Privileged

Ratio of the FNR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false negative rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

21. False Omission Rate Ratio

False Omission Rate Ratio = FOR Unprivileged / FOR Privileged

Ratio of the FOR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false omission rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

22. False Discovery Rate Ratio

False Discovery Rate Ratio = FDR Unprivileged / FDR Privileged

Ratio of the FDR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false discovery rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

23. Average Odds Difference

Average Odds Difference=((FPR unprivileged – FPR privileged) + (TPR unprivileged – TPR privileged)) * 0.5.

Average odds difference refers to the difference in probabilities or likelihoods between two groups or conditions. It is a measure used to compare the average chances or odds of an event occurring in one group compared to another.

Let's say we have two groups of people: Group A and Group B. We want to compare the average odds of a certain event happening in each group. If the average odds difference is positive, it means that the event is more likely to occur in Group A compared to Group B. On the other hand, if the average odds difference is negative, it means that the event is more likely to occur in Group B compared to Group A.

For example, let's consider the average odds difference of smoking-related diseases between smokers and non-smokers. If the average odds difference is positive, it indicates that smokers have a higher average likelihood of developing smoking-related diseases compared to non-smokers. Conversely, if the average odds difference is negative, it means that non-smokers have a higher average likelihood of avoiding smoking-related diseases compared to smokers.

24. Average Absolute Odds Difference

Average Absolute Odds Difference = (abs (FPR unprivileged – FPR Privileged) + abs(TPR unprivileged – TPR Privileged)) * 0.5

Average Absolute Odds Difference is a measure used to compare the effectiveness of two different models or treatments. It quantifies the difference in the likelihood of an event occurring between the two models or treatments.

In simpler terms, let's say we have two models that predict whether it will rain tomorrow. Model A predicts a 70% chance of rain, while Model B predicts a 60% chance of rain. The difference between these two predictions is 10%. The Average Absolute Odds Difference takes into account the absolute value of this difference, meaning it ignores whether it is positive or negative.

By calculating the average of these absolute differences across multiple predictions or scenarios, we can get a sense of how much the two models or treatments differ in their predictions or outcomes.

25. Error Rate Difference

Error Rate Difference = Error rate (Unprivileged) – Error Rate (Privileged)

Difference between error rates of Privileged and unprivileged group.

The ideal value of this metric is 0. A value of < 0 implies higher benefit for the unprivileged group and a value > 0 implies higher benefit for the privileged group.

26. Error Rate Ratio

Error Rate Ratio = Error rate (Unprivileged) / Error Rate (Privileged)

Ratio between error rates of Privileged and unprivileged group

The ideal value of this metric is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the privileged group.

27. Selection Rate

Selection Rate = (Fp+TP)/ Total_no_of_points

Out of all points, how many of them are predicted as positive.

28. Disparate Impact Ratio

Disparate Impact Ratio =
No_Predicted_positive_Unprivileged_Group / No_Predicted_positive_Privileged_Group.

Disparate Impact Ratio refers to a measure used to assess whether a particular policy or decision has a disproportionately negative impact on a certain group of people, based on their protected characteristics such as race, gender, or age.

In simpler terms, it looks at whether a rule or practice unintentionally affects one group more negatively than another. For example, if a company has a hiring policy that results in a significantly lower percentage of women being hired compared to men, it may indicate a disparate impact.

The ideal value of this metric is 1.0 A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

29. Statistical Parity Difference

Statistical Parity Difference = No of Predicted positive Unprivileged Group - No of Predicted positive Privileged Group

Statistical Parity Difference, in simple terms, refers to a measure of the difference in outcomes between different groups of people. It is used to assess whether there is a fair and equal distribution of opportunities or outcomes across different groups.

Imagine you have two groups of people, Group A and Group B. Statistical Parity Difference looks at the difference in outcomes, such as employment rates or loan approvals, between these two groups. If the Statistical Parity Difference is zero, it means that both groups have the same outcomes and there is no disparity or bias.

A value < 0 implies a higher benefit for the privileged group and a value >0 implies a higher benefit for the unprivileged group.

30. Equal Opportunity Difference

This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.

Equal opportunity means that everyone has the same chances and opportunities, regardless of their background or characteristics.

It ensures that people are treated fairly and have an equal shot at success. The concept of equal opportunity recognizes that everyone should have an equal chance to pursue their goals and dreams, without facing discrimination or barriers based on factors like race, gender, or socioeconomic status.

The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.