Infosys
Responsible AI Office

Infosys Responsible AI Toolkit

Fairness & Bias

API usage Instructions

Contents

Introduction

The following set of endpoints helps in analyzing the fairness of both traditional and generative AI models. Endpoints associated with traditional models help detect, mitigate bias, and generate a detailed report. For generative AI models, endpoints identify and categorize associated biases.

Once API swagger page is populated as per instructions given in the github repository Readme file, click on 'try it out' to use required endpoints. Details of endpoints associated with Fairness repository are outlined below.

Fairness Analyze

This endpoint is used to analyze the pretrain data and post-train data [with model's predictions] for group bias using metrics like Statistical parity.

Endpoint: /api/v1/fairness/Analyse

Payload Details:

biasType: Provide bias type based on your requirement PRETRAIN/POSTTRAIN.

methodType: Provided method type for metric score like disparate impact or ALL will return available metric scores. **taskType**: As of now we have only CLASSIFICATION.

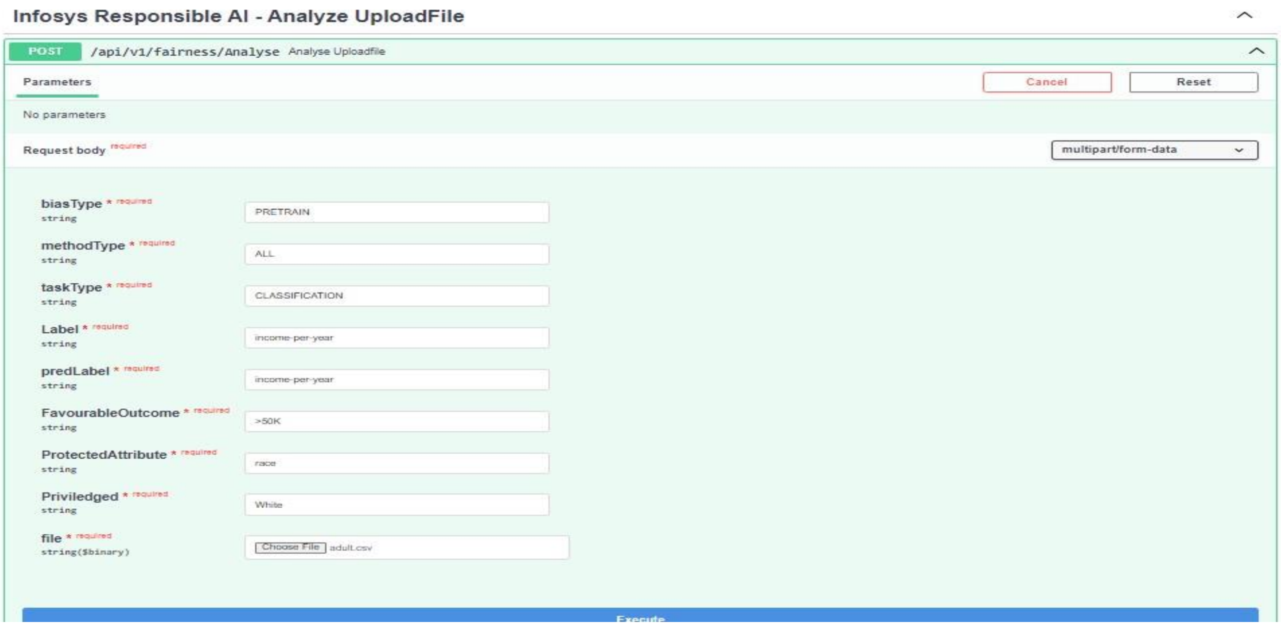
Label: Mention the target column name to predict.

predLabel: Add prediction label. Default is "labels_pred". This is required for POSTTRAIN.

FavourableOutcome: Mention favorable outcome for predict column.

ProtectedAttribute: Mention the protected attribute column name.

Privileged: Mention the privileged value for protected attribute. If multiple Privileged groups are there, entered in this format [priv_1,priv_2],[priv_3,priv_4] **File**: Upload the dataset.



Infosys Responsible AI - Analyze UploadFile

POST /api/v1/fairness/Analyse Analyse Uploadfile

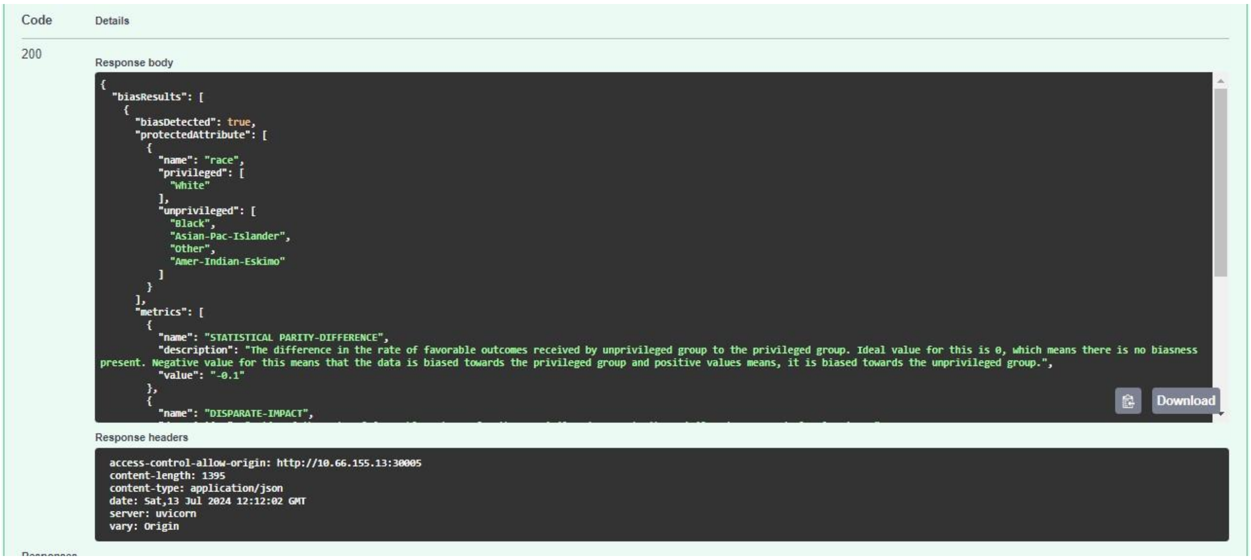
Parameters

No parameters

Request body required multipart/form-data

biasType * required string	PRETRAIN
methodType * required string	ALL
taskType * required string	CLASSIFICATION
Label * required string	income-per-year
predLabel * required string	income-per-year
FavourableOutcome * required string	>50K
ProtectedAttribute * required string	race
Priviledged * required string	White
file * required string($binary)	Choose File adult.csv

Response:



Code Details

200

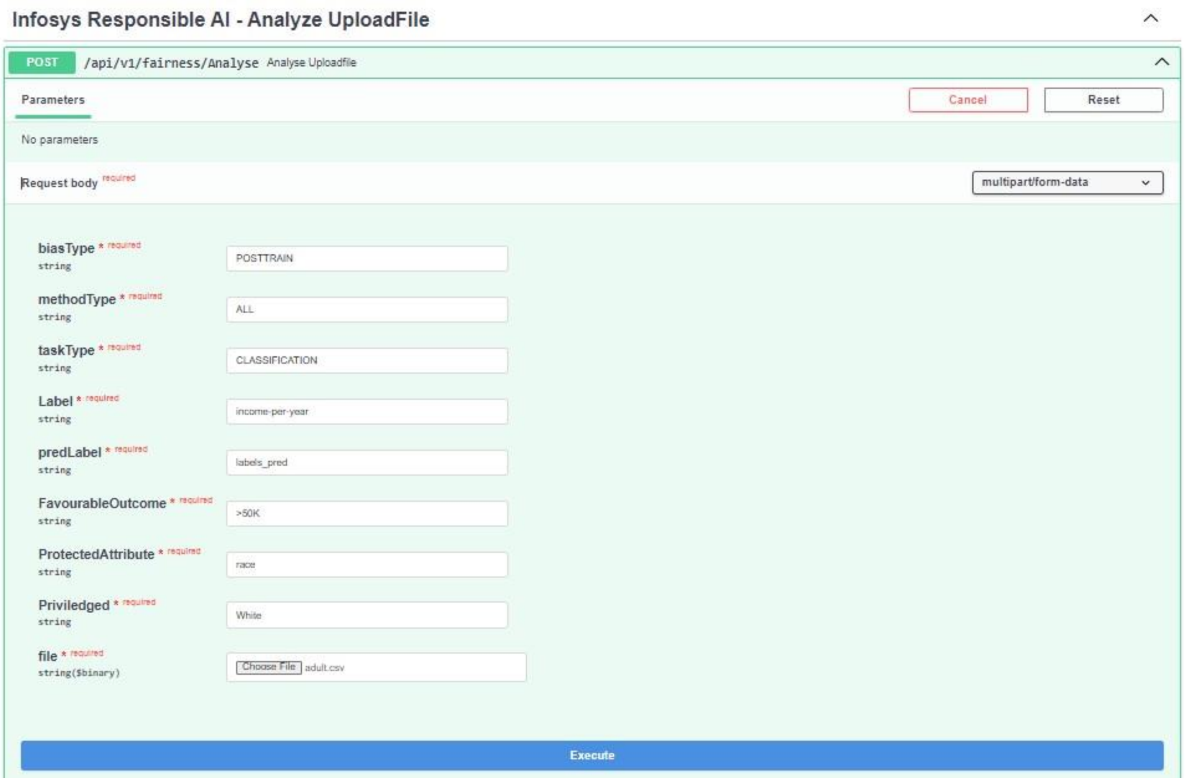
Response body

```
{
  "biasResults": [
    {
      "biasDetected": true,
      "protectedAttribute": [
        {
          "name": "race",
          "privileged": [
            "White"
          ],
          "unprivileged": [
            "Black",
            "Asian-Pac-Islander",
            "Other",
            "Amer-Indian-Eskimo"
          ]
        }
      ],
      "metrics": [
        {
          "name": "STATISTICAL PARITY-DIFFERENCE",
          "description": "The difference in the rate of favorable outcomes received by unprivileged group to the privileged group. Ideal value for this is 0, which means there is no biasness present. Negative value for this means that the data is biased towards the privileged group and positive values means, it is biased towards the unprivileged group.",
          "value": "-0.1"
        },
        {
          "name": "DISPARATE-IMPACT",
```

Response headers

```
access-control-allow-origin: http://10.66.155.13:30005
content-length: 1395
content-type: application/json
date: Sat,13 Jul 2024 12:12:02 GMT
server: uvicorn
vary: Origin
```

For Post Train



Infosys Responsible AI - Analyze UploadFile

POST /api/v1/fairness/Analyse Analyse Uploadfile

No parameters

Request body required multipart/form-data

biasType * required string: POSTTRAIN

methodType * required string: ALL

taskType * required string: CLASSIFICATION

Label * required string: income-per-year

predLabel * required string: labels_pred

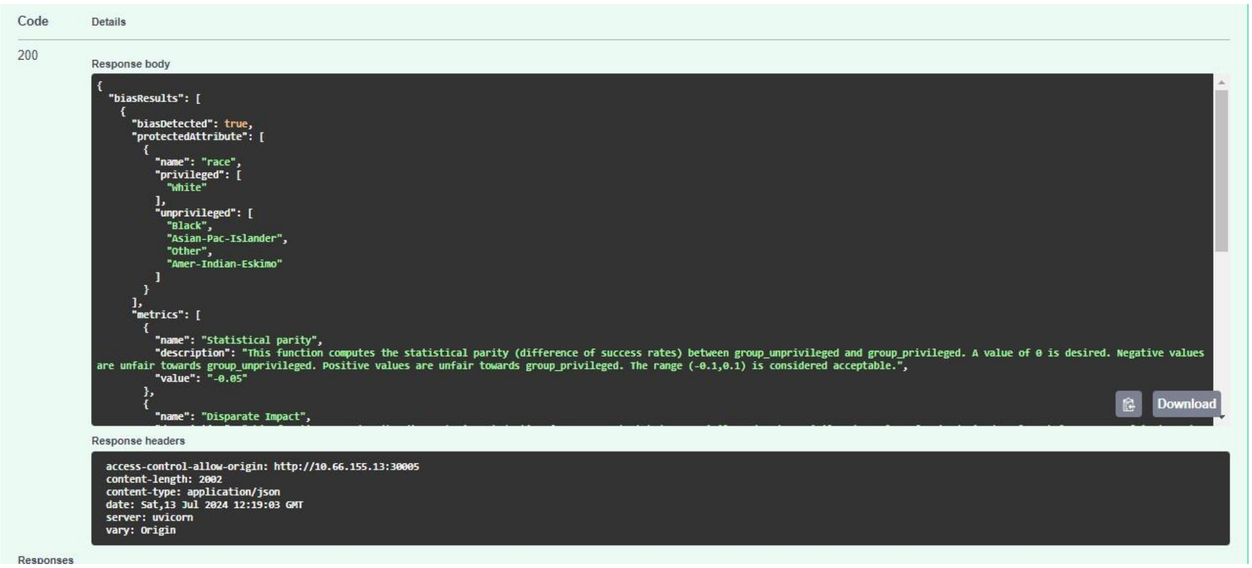
FavourableOutcome * required string: >50K

ProtectedAttribute * required string: race

Priviledged * required string: White

file * required string($binary): Choose File adult.csv

Response:



200

Response body

```
{
  "biasResults": [
    {
      "biasDetected": true,
      "protectedAttribute": [
        {
          "name": "race",
          "privileged": [
            "White"
          ],
          "unprivileged": [
            "Black",
            "Asian-Pac-Islander",
            "Other",
            "Amer-Indian-Eskimo"
          ]
        }
      ],
      "metrics": [
        {
          "name": "Statistical parity",
          "description": "This function computes the statistical parity (difference of success rates) between group_unprivileged and group_privileged. A value of 0 is desired. Negative values are unfair towards group_unprivileged. Positive values are unfair towards group_privileged. The range (-0.1,0.1) is considered acceptable.",
          "value": "-0.05"
        },
        {
          "name": "Disparate Impact",
```

Response headers

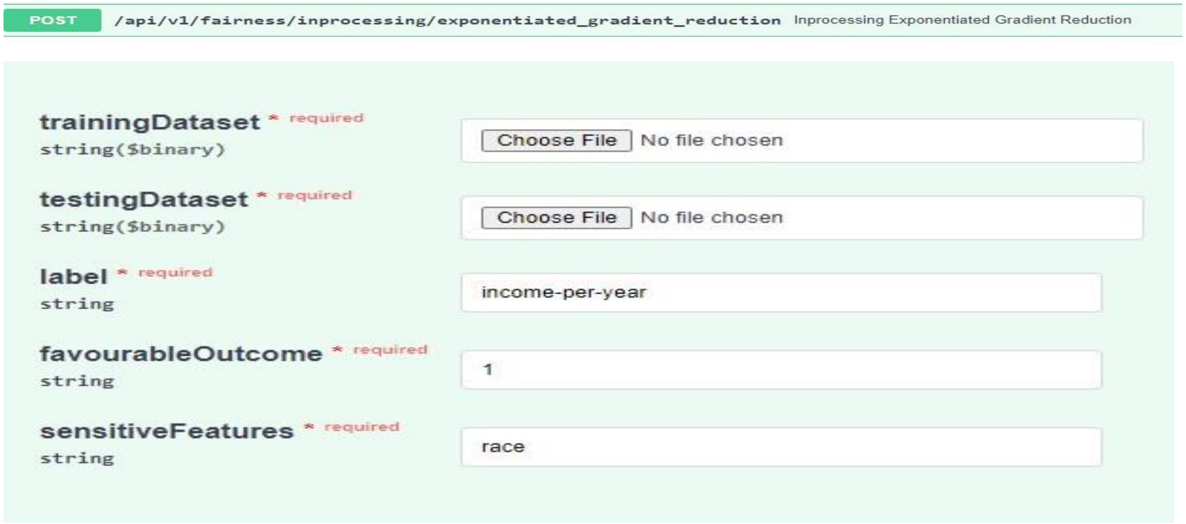
```
access-control-allow-origin: http://10.66.155.13:30005
content-length: 2002
content-type: application/json
date: Sat,13 Jul 2024 12:19:03 GMT
server: uvicorn
vary: Origin
```

In-Processing Analyze

This endpoint is used to instantiate a binary classification model and train with the train dataset uploaded along with the information of the sensitive columns in the dataset. The trained model would be aware of the sensitive attributes.

Endpoint: api/v1/fairness/inprocessing/exponentiated_gradient_reduction

Step1: Please go to the "exponentiated_gradient_reduction" API at the URL mentioned above.



POST
/api/v1/fairness/inprocessing/exponentiated_gradient_reduction
Inprocessing Exponentiated Gradient Reduction
trainingDataset * required
string($binary)
Choose File
No file chosen
testingDataset * required
string($binary)
Choose File
No file chosen
label * required
string
favourableOutcome * required
string
sensitiveFeatures * required
string

label: income-per-year

favourableOutcome: 1

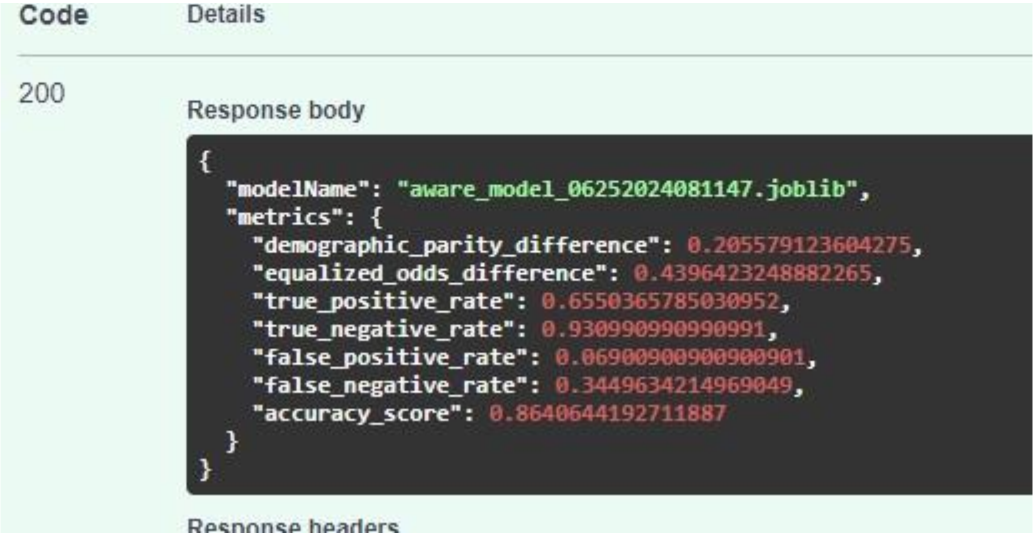
sensitiveFeatures: race

Label: Mention the target column name to predict.

FavourableOutcome: Mention favourable outcome for predict column.

ProtectedAttribute: Mention the protected attribute column name.

Provide datasets required to execute.



Code Details

200

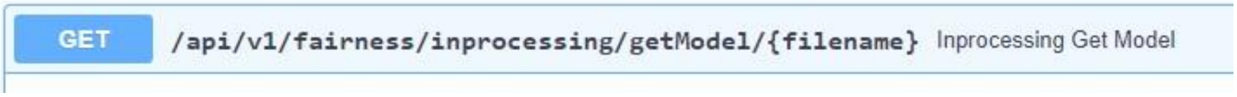
Response body

```
{
  "modelName": "aware_model_06252024081147.joblib",
  "metrics": {
    "demographic_parity_difference": 0.205579123604275,
    "equalized_odds_difference": 0.4396423248882265,
    "true_positive_rate": 0.6550365785030952,
    "true_negative_rate": 0.930990990990991,
    "false_positive_rate": 0.06900900900900901,
    "false_negative_rate": 0.3449634214969049,
    "accuracy_score": 0.8640644192711887
  }
}
```

Response headers

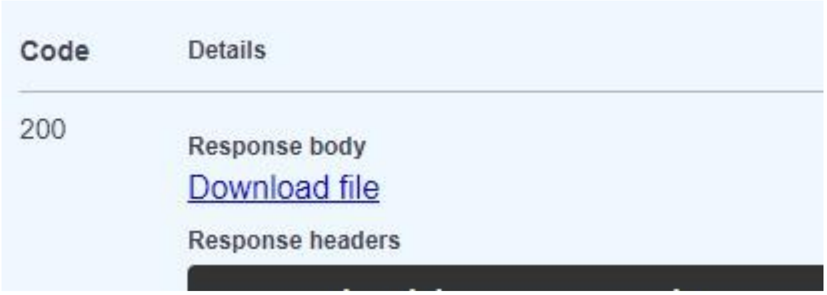
You will receive the return of metric scores, return optimized model name as a response.

Step2: Please go to the "getModel/{filename}" API at the URL mentioned above.



GET /api/v1/fairness/inprocessing/getModel/{filename} Inprocessing Get Model

Provide the filename to download the model.



Code Details

200

Response body

Download file

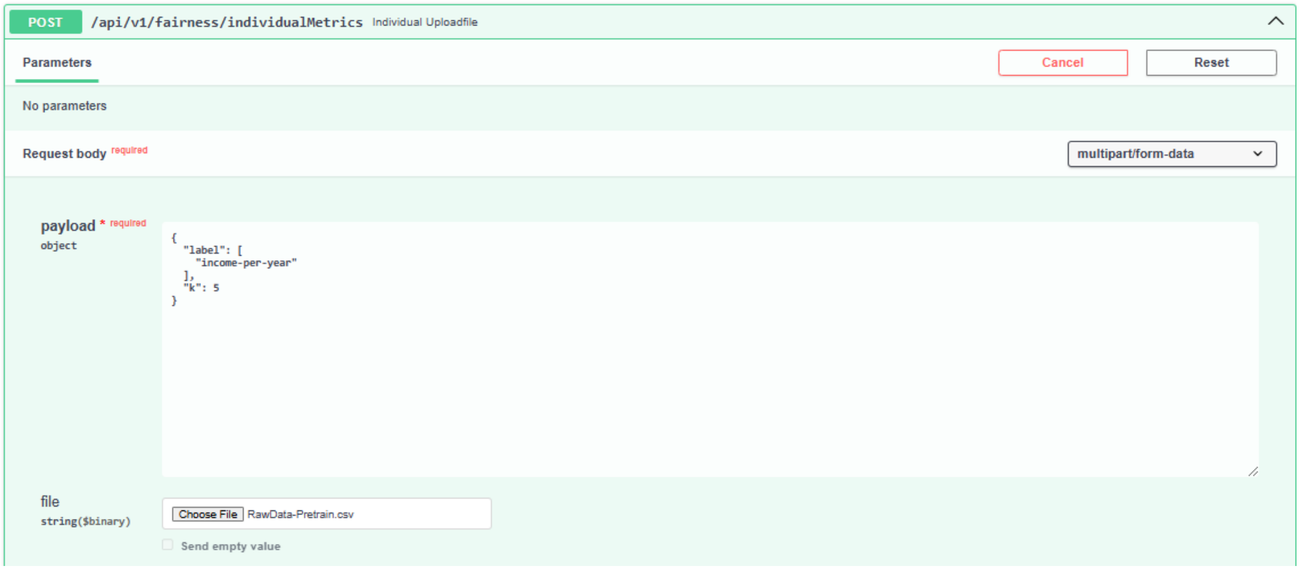
Response headers

Click on download file to save into local.

Individual metrics:

This endpoint is used to analyze the pretrain data and post-train data [with model's predictions] for group bias using metrics like Statistical parity.

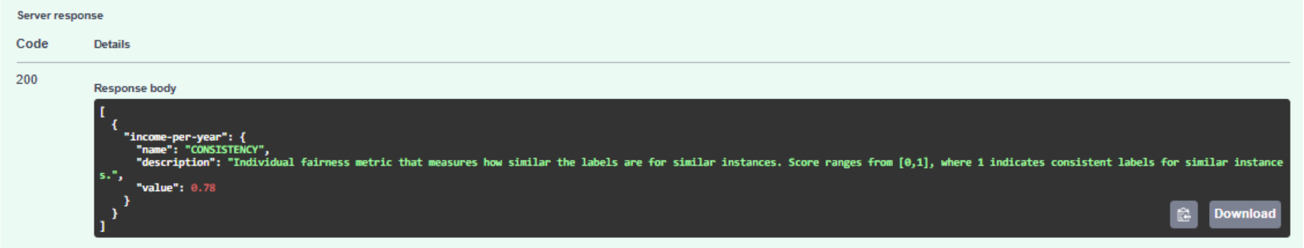
Endpoint: /api/v1/fairness/individualMetrics



```
{
  "label": [
    "income-per-year"
  ],
  "k": 5
}
```

RawData-Pretrain.csv

Upload the file to return attributes for dataset. Provide value for k as 5 and the label column where ground truth is available.

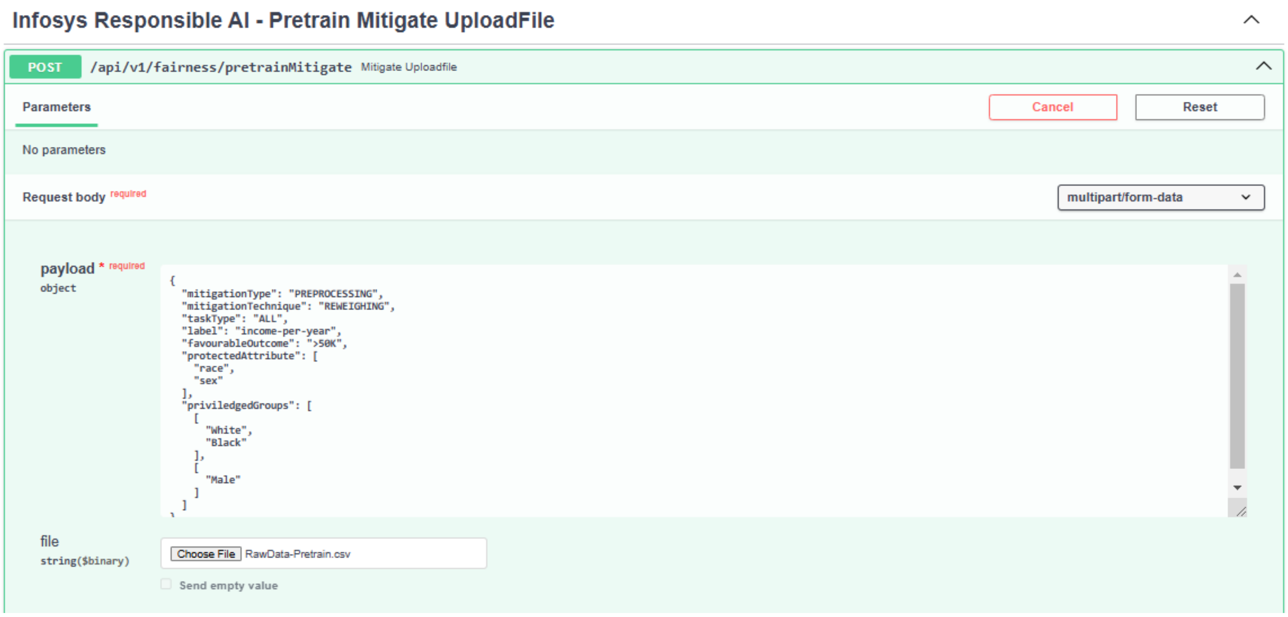


```
[
  {
    "income-per-year": {
      "name": "CONSISTENCY",
      "description": "Individual fairness metric that measures how similar the labels are for similar instances. Score ranges from [0,1], where 1 indicates consistent labels for similar instance
s.",
      "value": 0.78
    }
  }
]
```

You will receive the metrics score for provided dataset.

Mitigate Dataset

Endpoint: api/v1/fairness/pretrain/mitigation/getDataset



Infosys Responsible AI - Pretrain Mitigate UploadFile

POST /api/v1/fairness/pretrainMitigate Mitigate Uploadfile

Parameters

No parameters

Request body required — multipart/form-data

payload * required
object

```
{
  "mitigationType": "PREPROCESSING",
  "mitigationTechnique": "REWEIGHING",
  "taskType": "ALL",
  "label": "income-per-year",
  "favourableOutcome": ">50K",
  "protectedAttribute": [
    "race",
    "sex"
  ],
  "priviledgedGroups": [
    [
      "White",
      "Black"
    ],
    [
      "Male"
    ]
  ]
}
```

file
string($binary)

Choose File RawData-Pretrain.csv

□ Send empty value

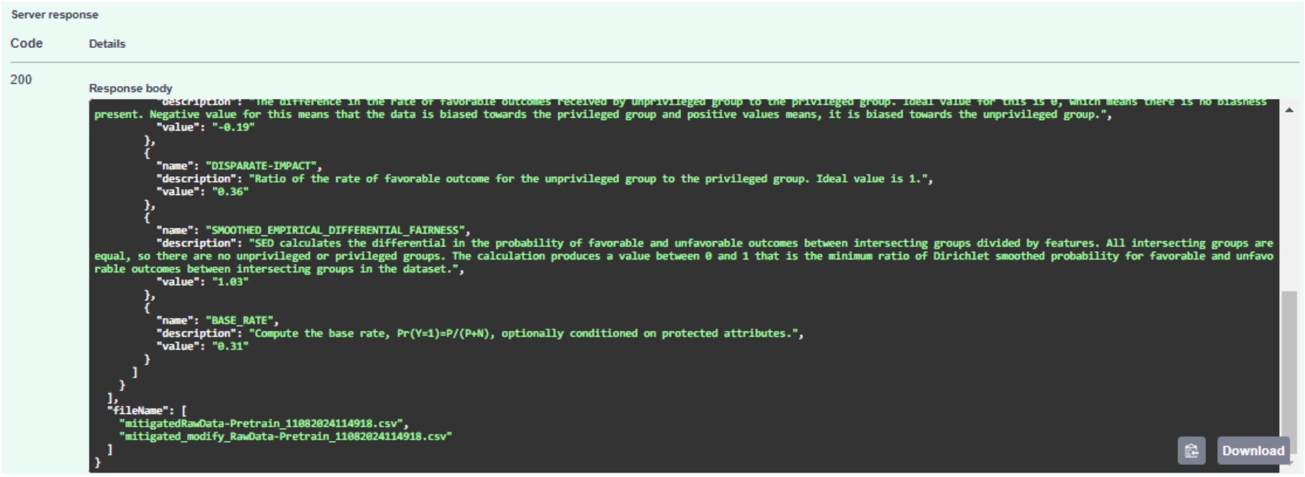
MitigationType: Mention the mitigationType Preprocessing.
MitigationTechinque: Mention the mitigationTechinque to mitigate.
taskType: As of now we have only CLASSIFICATION.
Label: Mention the target column name to predict.

FavourableOutcome: Mention favourable outcome for predict column.

ProtectedAttribute: Mention the protected attribute column name.

Privileged: Mention the priviledged value for protected attribute.

You will receive the metrics score for provided dataset and mitigated filename as a response.



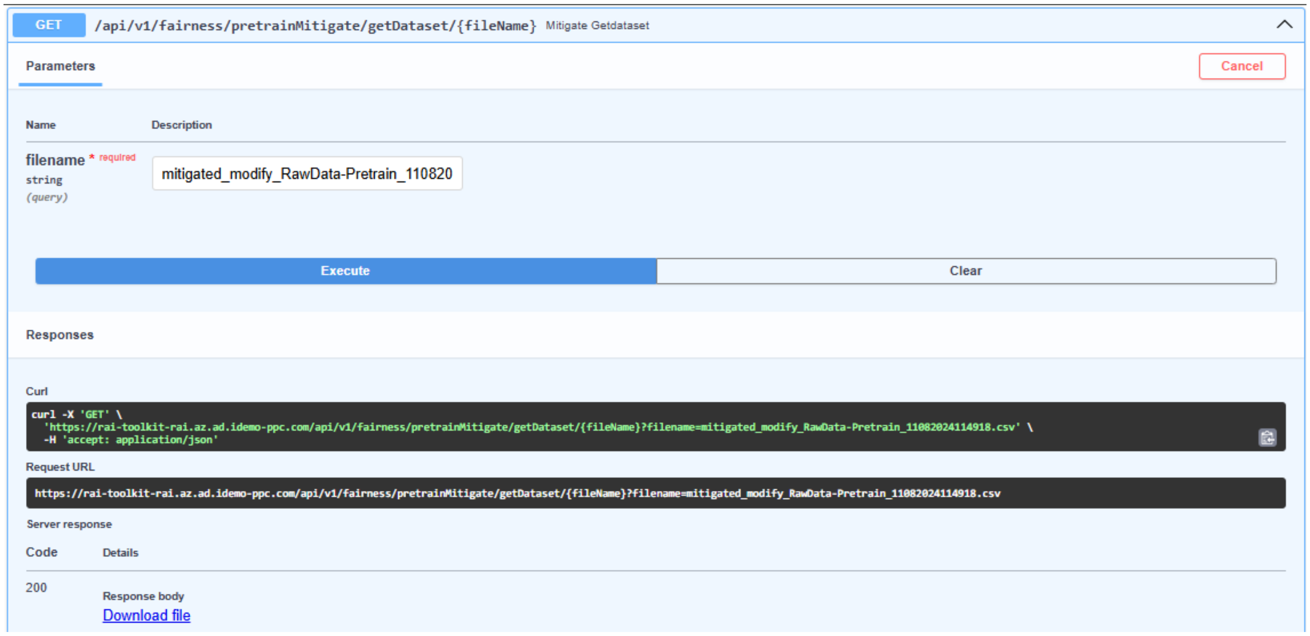
Server response

Code: 200

Details — Response body

```
          "description": "The difference in the rate of favorable outcomes received by unprivileged group to the privileged group. Ideal value for this is 0, which means there is no biasness present. Negative value for this means that the data is biased towards the privileged group and positive values means, it is biased towards the unprivileged group.",
          "value": "-0.19"
        },
        {
          "name": "DISPARATE-IMPACT",
          "description": "Ratio of the rate of favorable outcome for the unprivileged group to the privileged group. Ideal value is 1.",
          "value": "0.36"
        },
        {
          "name": "SMOOTHED_EMPIRICAL_DIFFERENTIAL_FAIRNESS",
          "description": "SED calculates the differential in the probability of favorable and unfavorable outcomes between intersecting groups divided by features. All intersecting groups are equal, so there are no unprivileged or privileged groups. The calculation produces a value between 0 and 1 that is the minimum ratio of Dirichlet smoothed probability for favorable and unfavorable outcomes between intersecting groups in the dataset.",
          "value": "1.03"
        },
        {
          "name": "BASE_RATE",
          "description": "Compute the base rate, Pr(Y=1)=P/(P+N), optionally conditioned on protected attributes.",
          "value": "0.31"
        }
      ]
    }
  ],
  "fileName": [
    "mitigatedRawData-Pretrain_11082024114918.csv",
    "mitigated_modify_RawData-Pretrain_11082024114918.csv"
  ]
}
```

Use above endpoint to download mitigate dataset by provided filename.



filename: mitigated_modify_RawData-Pretrain_110820

Curl

```
curl -X 'GET' \
  'https://rai-toolkit-rai.az.ad.idemo-ppc.com/api/v1/fairness/pretrainMitigate/getDataset/{fileName}?filename=mitigated_modify_RawData-Pretrain_11082024114918.csv' \
  -H 'accept: application/json'
```

Request URL

```
https://rai-toolkit-rai.az.ad.idemo-ppc.com/api/v1/fairness/pretrainMitigate/getDataset/{fileName}?filename=mitigated_modify_RawData-Pretrain_11082024114918.csv
```

Server response

Code	Details
200	Response body Download file

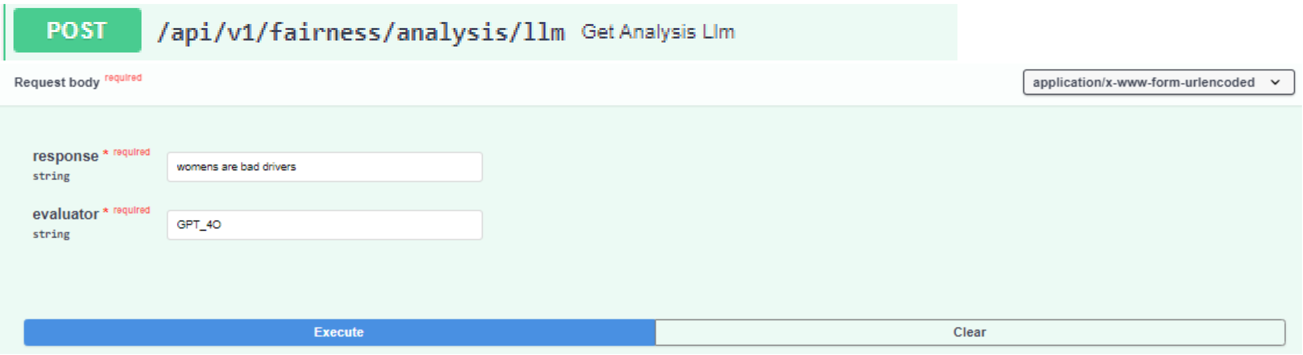
Click on download file to download the into local.

Generative AI

Text analysis using GPT

Step1:

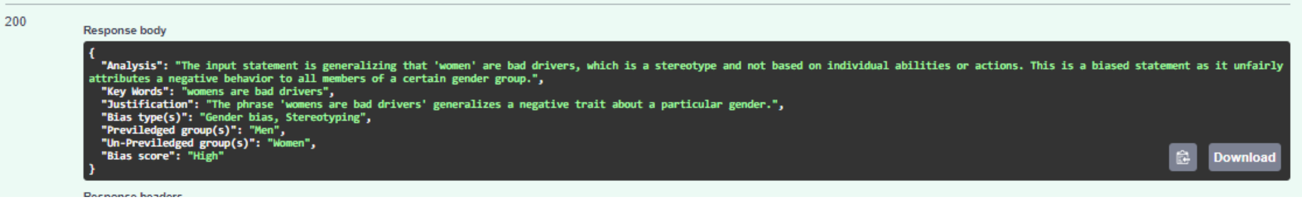
Endpoint: **/api/v1/fairness/analysis/llm**



response: womens are bad drivers

evaluator: GPT_4O

You will get response of bias analysis in response body.

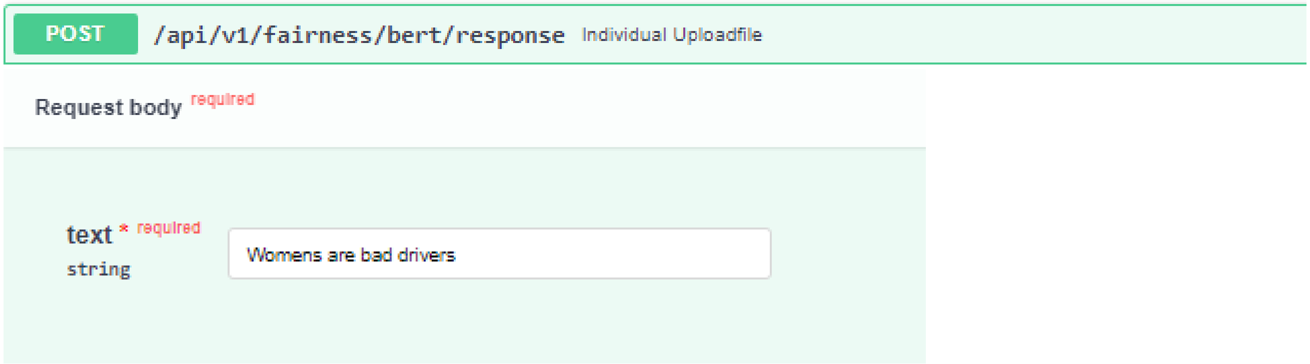


```
{
  "Analysis": "The input statement is generalizing that 'women' are bad drivers, which is a stereotype and not based on individual abilities or actions. This is a biased statement as it unfairly attributes a negative behavior to all members of a certain gender group.",
  "Key Words": "womens are bad drivers",
  "Justification": "The phrase 'womens are bad drivers' generalizes a negative trait about a particular gender.",
  "Bias type(s)": "Gender bias, Stereotyping",
  "Previledged group(s)": "Men",
  "Un-Previledged group(s)": "Women",
  "Bias score": "High"
}
```

Text analysis using Bert

Step1:

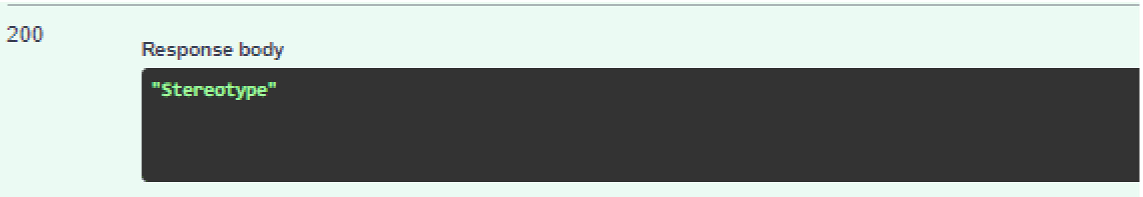
Endpoint: **/api/v1/fairness/bert/response**



text * required
string

Womens are bad drivers

You will get response of bias analysis in response body.

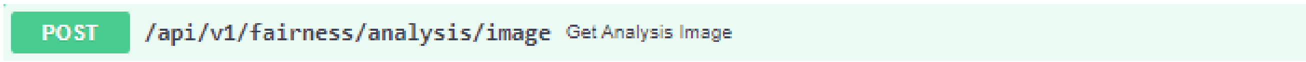


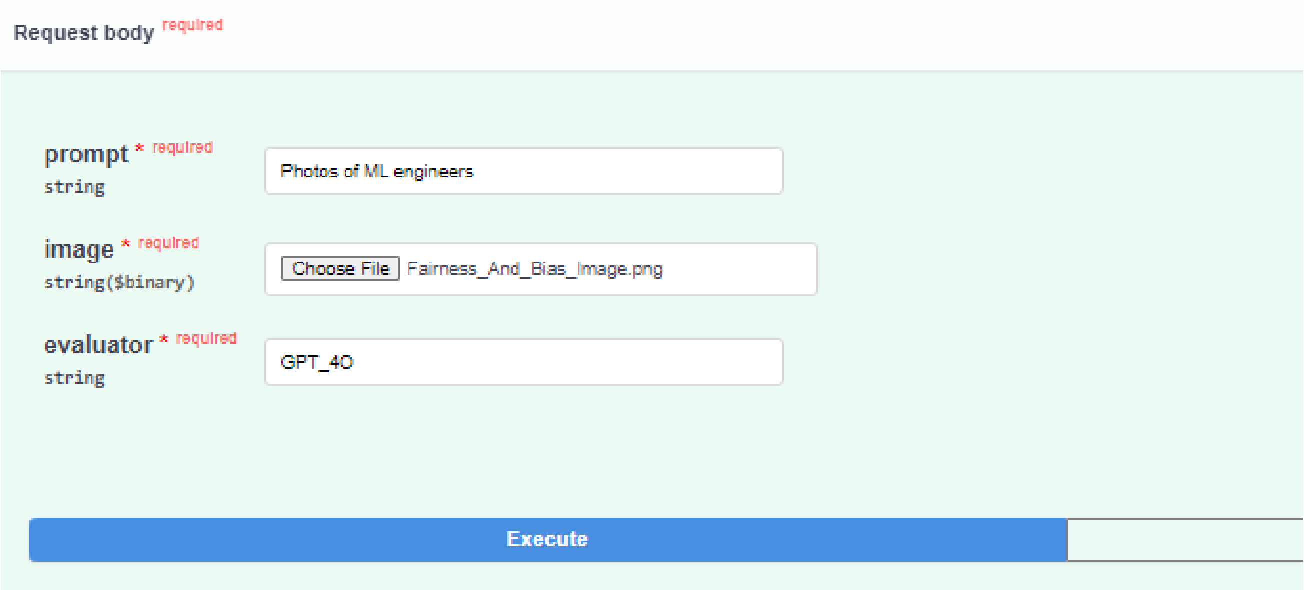
```
"Stereotype"
```

Image Analysis

Step1:

Endpoint: **/api/v1/fairness/analysis/image**





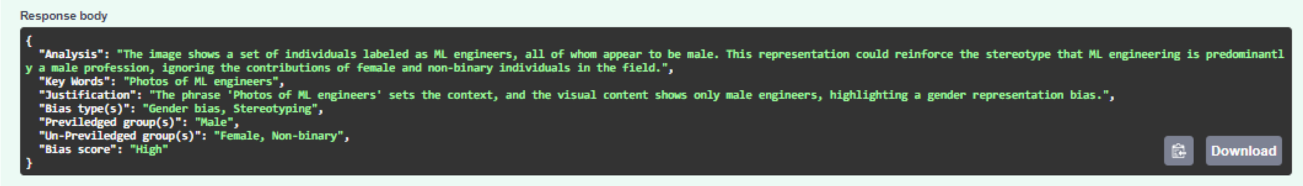
Request body required

prompt * required
string
Photos of ML engineers

image * required
string($binary)
Choose File Fairness_And_Bias_Image.png

evaluator * required
string
GPT_4O

You will get response of bias analysis in response body.



Response body

```
{
  "Analysis": "The image shows a set of individuals labeled as ML engineers, all of whom appear to be male. This representation could reinforce the stereotype that ML engineering is predominantl
y a male profession, ignoring the contributions of female and non-binary individuals in the field.",
  "Key Words": "Photos of ML engineers",
  "Justification": "The phrase 'Photos of ML engineers' sets the context, and the visual content shows only male engineers, highlighting a gender representation bias.",
  "Bias type(s)": "Gender bias, Stereotyping",
  "Previledged group(s)": "Male",
  "Un-Previledged group(s)": "Female, Non-binary",
  "Bias score": "High"
}
```