Installing required libraries

In [1]:
```
pip install aif360
```

```
Collecting aif360
  Downloading aif360-0.5.0-py3-none-any.whl (214 kB)
     |████████████████████████████████| 214 kB 2.4 MB/s eta 0:00:01
Requirement already satisfied: matplotlib in ./opt/anaconda3/lib/python3.9/site-packages
(from aif360) (3.5.1)
Requirement already satisfied: pandas>=0.24.0 in ./opt/anaconda3/lib/python3.9/site-pack
ages (from aif360) (1.4.2)
Requirement already satisfied: numpy>=1.16 in ./opt/anaconda3/lib/python3.9/site-package
s (from aif360) (1.21.5)
Requirement already satisfied: scipy>=1.2.0 in ./opt/anaconda3/lib/python3.9/site-packag
es (from aif360) (1.7.3)
Requirement already satisfied: scikit-learn>=1.0 in ./opt/anaconda3/lib/python3.9/site-p
ackages (from aif360) (1.0.2)
Requirement already satisfied: python-dateutil>=2.8.1 in ./opt/anaconda3/lib/python3.9/s
ite-packages (from pandas>=0.24.0->aif360) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in ./opt/anaconda3/lib/python3.9/site-packag
es (from pandas>=0.24.0->aif360) (2021.3)
Requirement already satisfied: six>=1.5 in ./opt/anaconda3/lib/python3.9/site-packages
(from python-dateutil>=2.8.1->pandas>=0.24.0->aif360) (1.16.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in ./opt/anaconda3/lib/python3.9/sit
e-packages (from scikit-learn>=1.0->aif360) (2.2.0)
Requirement already satisfied: joblib>=0.11 in ./opt/anaconda3/lib/python3.9/site-packag
es (from scikit-learn>=1.0->aif360) (1.1.0)
Requirement already satisfied: pillow>=6.2.0 in ./opt/anaconda3/lib/python3.9/site-packa
ges (from matplotlib->aif360) (9.0.1)
Requirement already satisfied: kiwisolver>=1.0.1 in ./opt/anaconda3/lib/python3.9/site-p
ackages (from matplotlib->aif360) (1.3.2)
Requirement already satisfied: packaging>=20.0 in ./opt/anaconda3/lib/python3.9/site-pac
kages (from matplotlib->aif360) (21.3)
Requirement already satisfied: cycler>=0.10 in ./opt/anaconda3/lib/python3.9/site-packag
es (from matplotlib->aif360) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in ./opt/anaconda3/lib/python3.9/site-p
ackages (from matplotlib->aif360) (4.25.0)
Requirement already satisfied: pyparsing>=2.2.1 in ./opt/anaconda3/lib/python3.9/site-pa
ckages (from matplotlib->aif360) (3.0.4)
Installing collected packages: aif360
Successfully installed aif360-0.5.0
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
pip install fairlearn
```

```
Collecting fairlearn
  Downloading fairlearn-0.8.0-py3-none-any.whl (235 kB)
     |████████████████████████████████| 235 kB 5.8 MB/s eta 0:00:01
Requirement already satisfied: scikit-learn>=0.22.1 in ./opt/anaconda3/lib/python3.9/sit
e-packages (from fairlearn) (1.0.2)
Requirement already satisfied: scipy>=1.4.1 in ./opt/anaconda3/lib/python3.9/site-packag
es (from fairlearn) (1.7.3)
Requirement already satisfied: numpy>=1.17.2 in ./opt/anaconda3/lib/python3.9/site-packa
ges (from fairlearn) (1.21.5)
Requirement already satisfied: pandas>=0.25.1 in ./opt/anaconda3/lib/python3.9/site-pack
ages (from fairlearn) (1.4.2)
Requirement already satisfied: python-dateutil>=2.8.1 in ./opt/anaconda3/lib/python3.9/s
ite-packages (from pandas>=0.25.1->fairlearn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in ./opt/anaconda3/lib/python3.9/site-packag
es (from pandas>=0.25.1->fairlearn) (2021.3)
Requirement already satisfied: six>=1.5 in ./opt/anaconda3/lib/python3.9/site-packages
(from python-dateutil>=2.8.1->pandas>=0.25.1->fairlearn) (1.16.0)
Requirement already satisfied: joblib>=0.11 in ./opt/anaconda3/lib/python3.9/site-packag
```

```
es (from scikit-learn>=0.22.1->fairlearn) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in ./opt/anaconda3/lib/python3.9/sit
e-packages (from scikit-learn>=0.22.1->fairlearn) (2.2.0)
Installing collected packages: fairlearn
Successfully installed fairlearn-0.8.0
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
import numpy as np
from aif360.datasets import GermanDataset
from aif360.metrics import BinaryLabelDatasetMetric
from aif360.algorithms.preprocessing import Reweighing
```

```
WARNING:root:No module named 'tempeh': LawSchoolGPADataset will be unavailable. To insta
ll, run:
pip install 'aif360[LawSchoolGPA]'
WARNING:root:No module named 'tensorflow': AdversarialDebiasing will be unavailable. To
install, run:
pip install 'aif360[AdversarialDebiasing]'
WARNING:root:No module named 'tensorflow': AdversarialDebiasing will be unavailable. To
install, run:
pip install 'aif360[AdversarialDebiasing]'
```

Loading the data

In [47]:
```
dataset_orig = GermanDataset(
    protected_attribute_names=['age'],
    privileged_classes=[lambda x: x >= 25],
    features_to_drop=['personal_status', 'sex']
   )
```

In [43]:
```
dataset_orig
```

Out[43]:

```
               instance weights features                \

                                    month credit_amount
instance names
0                             1.0      6.0        1169.0
1                             1.0     48.0        5951.0
2                             1.0     12.0        2096.0
3                             1.0     42.0        7882.0
4                             1.0     24.0        4870.0
...                           ...      ...           ...
995                           1.0     12.0        1736.0
996                           1.0     30.0        3857.0
997                           1.0     12.0         804.0
998                           1.0     45.0        1845.0
999                           1.0     45.0        4576.0

                                                        \

               investment_as_income_percentage residence_since
instance names
0                                          4.0             4.0
1                                          2.0             2.0
2                                          2.0             3.0
3                                          2.0             4.0
4                                          3.0             4.0
...                                        ...             ...
995                                        3.0             4.0
996                                        4.0             4.0
997                                        4.0             4.0
998                                        4.0             4.0
999                                        3.0             4.0
```

```
                                                                        \
               protected attribute
                                age number_of_credits people_liable_for
instance names
0                                1.0               2.0               1.0
1                                0.0               1.0               1.0
2                                1.0               1.0               2.0
3                                1.0               1.0               2.0
4                                1.0               2.0               2.0
...                              ...               ...               ...
995                              1.0               1.0               1.0
996                              1.0               1.0               1.0
997                              1.0               1.0               1.0
998                              0.0               1.0               1.0
999                              1.0               1.0               1.0

                                          ...                            \
                                          ...
               status=A11 status=A12  ... housing=A153 skill_level=A171
instance names                            ...
0                     1.0        0.0  ...          0.0               0.0
1                     0.0        1.0  ...          0.0               0.0
2                     0.0        0.0  ...          0.0               0.0
3                     1.0        0.0  ...          1.0               0.0
4                     1.0        0.0  ...          1.0               0.0
...                   ...        ...  ...          ...               ...
995                   0.0        0.0  ...          0.0               0.0
996                   1.0        0.0  ...          0.0               0.0
997                   0.0        0.0  ...          0.0               0.0
998                   1.0        0.0  ...          1.0               0.0
999                   0.0        1.0  ...          0.0               0.0

                                                                       \

               skill_level=A172 skill_level=A173 skill_level=A174
instance names
0                           0.0              1.0               0.0
1                           0.0              1.0               0.0
2                           1.0              0.0               0.0
3                           0.0              1.0               0.0
4                           0.0              1.0               0.0
...                         ...              ...               ...
995                         1.0              0.0               0.0
996                         0.0              0.0               1.0
997                         0.0              1.0               0.0
998                         0.0              1.0               0.0
999                         0.0              1.0               0.0

                                                                     \

               telephone=A191 telephone=A192 foreign_worker=A201
instance names
0                         0.0            1.0                1.0
1                         1.0            0.0                1.0
2                         1.0            0.0                1.0
3                         1.0            0.0                1.0
4                         1.0            0.0                1.0
...                       ...            ...                ...
995                       1.0            0.0                1.0
996                       0.0            1.0                1.0
997                       1.0            0.0                1.0
998                       0.0            1.0                1.0
999                       1.0            0.0                1.0

                                labels
```

```
                foreign_worker=A202
instance names
0                               0.0    1.0
1                               0.0    2.0
2                               0.0    1.0
3                               0.0    1.0
4                               0.0    2.0
...                             ...    ...
995                             0.0    1.0
996                             0.0    1.0
997                             0.0    1.0
998                             0.0    2.0
999                             0.0    1.0

[1000 rows x 59 columns]
```

Split the data into train/test dataset (70%)

In [48]:
```
dataset_orig_train, dataset_orig_test = dataset_orig.split([0.7], shuffle=True)
```

Privileged and unprivileged groups segregation

In [49]:
```
privileged_groups = [{'age': 1}]
unprivileged_groups = [{'age': 0}]
```

Fairness metrics in original training dataset (BinaryLabelDatsetMetric)

In [50]:
```
metric_orig_train = BinaryLabelDatasetMetric(dataset_orig_train,
                                             unprivileged_groups=unprivileged_groups,
                                             privileged_groups=privileged_groups)

print("Difference in mean outcomes between unprivileged and privileged groups = %f" % me
```

```
Difference in mean outcomes between unprivileged and privileged groups = -0.117438
```

The difference in mean outcomes is -0.117 indicating that privileged data is almost 11.7% more than unprivileged

Mitigation of this bias of 11% using AI fairness

Reweighing the groups

In [51]:
```
RW = Reweighing(unprivileged_groups=unprivileged_groups,
                privileged_groups=privileged_groups)
dataset_transf_train = RW.fit_transform(dataset_orig_train)
```

Check the difference now

In [52]:
```
metric_transf_train = BinaryLabelDatasetMetric(dataset_transf_train,
                                               unprivileged_groups=unprivileged_groups,
                                               privileged_groups=privileged_groups)

print("Difference in mean outcomes between unprivileged and privileged groups = %f" % me
```

```
Difference in mean outcomes between unprivileged and privileged groups = 0.000000
```

The difference has come down to 0 indicating very effective mitigation

Trying it out with the feature 'sex'

In [64]:
```
dataset_orig = GermanDataset(
    protected_attribute_names=['sex'],
    privileged_classes=[lambda x: x == 'male'],
    features_to_drop=['personal_status', 'age']
   )
```

In [58]:
```
dataset_orig_train, dataset_orig_test = dataset_orig.split([0.7], shuffle=True)
```

In [59]:
```
privileged_groups = [{'sex': 1}]
unprivileged_groups = [{'sex': 0}]
```

In [60]:
```
metric_orig_train = BinaryLabelDatasetMetric(dataset_orig_train,
                                             unprivileged_groups=unprivileged_groups,
                                             privileged_groups=privileged_groups)

print("Difference in mean outcomes between unprivileged and privileged groups = %f" % me
```

```
Difference in mean outcomes between unprivileged and privileged groups = -0.073420
```

In [61]:
```
RW = Reweighing(unprivileged_groups=unprivileged_groups,
                privileged_groups=privileged_groups)
dataset_transf_train = RW.fit_transform(dataset_orig_train)
```

In [62]:
```
metric_transf_train = BinaryLabelDatasetMetric(dataset_transf_train,
                                               unprivileged_groups=unprivileged_groups,
                                               privileged_groups=privileged_groups)

print("Difference in mean outcomes between unprivileged and privileged groups = %f" % me
```

```
Difference in mean outcomes between unprivileged and privileged groups = -0.000000
```

Summary

The AIFairness toolkitit is an open-source toolkit designed to help examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. In this assignment, the GermanDataset from aif360 is used, and the protected attribute is defined as the "age" column, with the privileged class being all individuals aged 25 or older.

The dataset is split into test and train sets, with 70% of the original dataset being used as the train set and 30% as the test set. The fairness metric is computed by calculating the difference between the percentage of favorable results for the privileged and unprivileged classes.

The Reweighing algorithm is then applied to transform the dataset and achieve more equity in positive outcomes on the protected attribute for both privileged and unprivileged groups. The biasness is recalculated after this transformation.

Based on the results, it seems that there was a significant difference of almost 11% in the fairness metric before applying AI Fairness. However, after applying the Reweighing algorithm, this difference was brought down to zero, indicating that the algorithm was successful in mitigating the bias in the dataset.

I tried the same methodology for the feature sex wherein there was a difference of 7% in the fairness metric between male and female, which was mitigated with the Reweighing algorithm

In [ ]: