1. Fairlearn

Что делает:

Предоставляет методы для оценки и смягчения несправедливости в моделях машинного обучения.

Как работает:

Использует алгоритмы, такие как Exponentiated Gradient и Grid Search, чтобы минимизировать несправедливость по заданным метрикам (например, Demographic Parity или Equalized Odds).

Плюсы:

- Легко интегрируется с scikit-learn.
- Гибкая настройка метрик справедливости.
- Визуализация в виде fairlearn.dashboard.

Минусы:

- Требует настройки sensitive features.
- Влияет на точность, поскольку жертвует ею ради справедливости.

2. AI Fairness 360 (AIF360)

Что делает:

Комплексный фреймворк от IBM для оценки и смягчения несправедливости на разных этапах ML-пайплайна.

Как работает:

Предлагает предобработку (например, Reweighing), алгоритмы обучения и постобработку для минимизации несправедливости.

Плюсы

- Поддержка множества метрик и методов.
- Может работать на любом этапе (до, во время и после обучения).
- Поддержка стандартных fairness-датасетов.

Минусы:

Более сложный в использовании, чем Fairlearn. Требует кастомной подготовки данных.

3. What-If Tool

Что делает: Визуальный инструмент от Google для анализа моделей без необходимости писать код.

Как работает: Позволяет интерактивно исследовать, как изменения в данных влияют на вывод модели.

Плюсы:

- Визуальный, понятный и интерактивный.
- Отлично подходит для объяснимости моделей.

Минусы:

- Ограничен в автоматизации и интеграции.
- Работает только с TensorBoard.

4. Themis-ML

Что делает:

Предлагает инструменты для оценки и устранения несправедливости в бинарных классификациях.

Как работает:

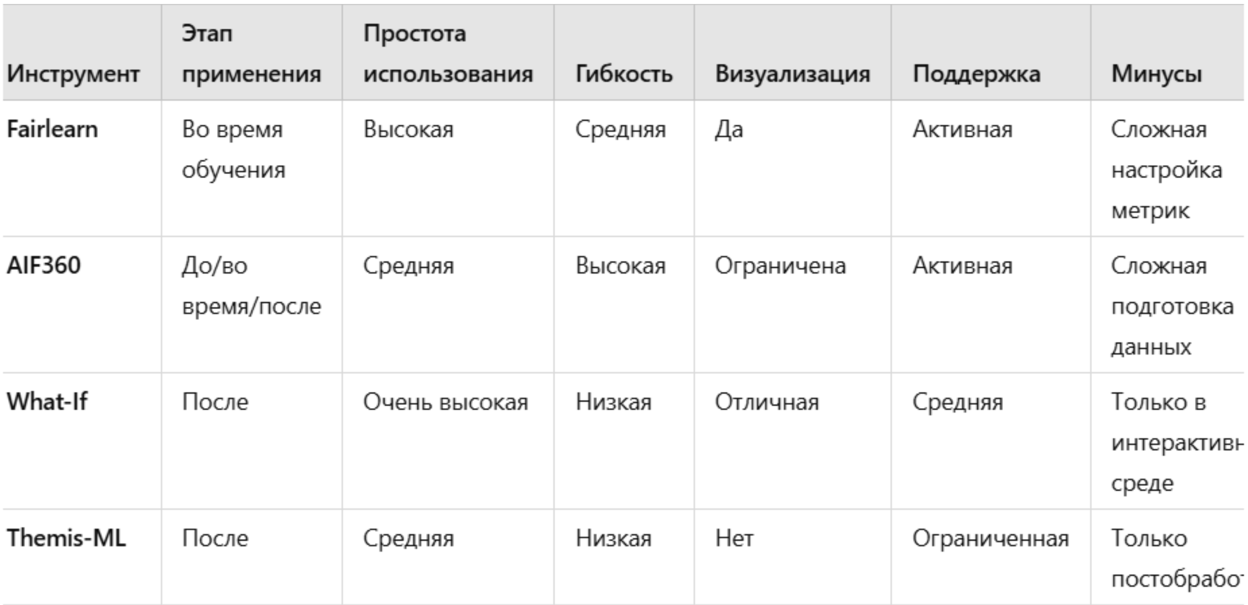
Использует методы постобработки, такие как Reject Option Classification, чтобы минимизировать несправедливость после обучения модели.

Плюсы:

- Простая реализация для постобработки.
- Работает с любыми классификационными моделями.

Минусы:

- Ограничен в возможностях (только постобработка).
- Меньше документации и поддержки.



Инструмент	Этап применения	Простота использования	Гибкость	Визуализация	Поддержка	Минусы
Fairlearn	Во время обучения	Высокая	Средняя	Да	Активная	Сложная настройка метрик
AIF360	До/во время/после	Средняя	Высокая	Ограничена	Активная	Сложная подготовка данных
What-If	После	Очень высокая	Низкая	Отличная	Средняя	Только в интерактивн среде
Themis-ML	После	Средняя	Низкая	Нет	Ограниченная	Только постобрабо

- Для быстрого прототипа — Fairlearn.

- Для глубокой работы на всех этапах — AIF360.
- Для интерактивного анализа — What-If Tool.
- Для простой постобработки — Themis-ML

Инструменты вроде Fairlearn, AI Fairness 360 и др. применяются для:

1. **Анализа справедливости моделей:**

есть проверка на предвзятость (bias) относительно чувствительных признаков (пол, раса и тд и тп) Применяется допустим при выдаче кредитов, наеме на работу, распределении ресурсов между пациентами в медицине(там, где может возникнуть предвзятость, соответсвенно неравное отношение к людям, исходящее из каких-либо факторов, типо пола, возраста, расы, национальности и тд)

1. **Корректировки смещений:**

Устранение предвзятости на этапах подготовки данных, обучения или после него

1. **Визуализации и интерпретации:**

Исследование поведения модели для разных групп(есть ли предвзятость)

1. **Соответствия нормативным требованиям:** Обеспечение прозрачности и отсутствия дискриминации

Они не предназначены для санитайзинга данных (очистки/предобработки), но помогают обнаружить предвзятость в данных и моделях Для санитайзинга лучше использовать инструменты типо Pandas, Scikit-learn, Great Expectations

Поняла так, что инструменты для анализа смещений нужны чтобы обеспечить справедливость моделей, но не могут заменить санитайзинг данных

Добавление **What-If Tool** (WIT) в код требует дополнительных шагов. Этот инструмент предоставляет интерактивный виджет для анализа модели, но он работает только с моделями, совместимыми с TensorFlow или scikit-learn.

Для работы с WIT:

Модель должна быть обернута в специальный формат. Данные должны быть подготовлены в виде списка списков (или массива NumPy). Так же, WIT может вызывать проблемы с зависимостями, такими как protobuf. Чтобы минимизировать конфликты, надо использовать его отдельно от других инструментов.

```
# Установка необходимых библиотек
!pip install fairlearn aif360 scikit-learn matplotlib pandas
protobuf==3.20.3 witwidget

# Импорт библиотек
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from fairlearn.metrics import demographic_parity_difference,
MetricFrame
from fairlearn.postprocessing import ThresholdOptimizer
from aif360.datasets import BinaryLabelDataset
from aif360.algorithms.preprocessing import Reweighing
from aif360.metrics import BinaryLabelDatasetMetric
import matplotlib.pyplot as plt
from witwidget.notebook.visualization import WitWidget,
WitConfigBuilder

# Загрузка датасета German Credit
url =
"https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/
german/german.data"
columns = ['status', 'duration', 'credit_history', 'purpose',
'amount', 'savings', 'employment',
           'installment_rate', 'personal_status', 'other_debtors',
'residence_since', 'property',
           'age', 'other_installments', 'housing', 'credits', 'job',
'people_liable', 'telephone',
           'foreign_worker', 'label']
data = pd.read_csv(url, names=columns, delimiter=' ')
data['label'] = data['label'].apply(lambda x: 1 if x == 1 else 0)  #
Преобразование меток в бинарные

# Разделение данных на признаки и целевую переменную
X = data.drop('label', axis=1)
y = data['label']

# Кодирование категориальных признаков
X = pd.get_dummies(X)

# Разделение данных на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)

# Чувствительный признак (например, возраст > 25 лет)
s_train = (X_train['age'] > 25).astype(int)  # Чувствительный признак
для обучающей выборки
s_test = (X_test['age'] > 25).astype(int)    # Чувствительный признак
для тестовой выборки

# Обучение базовой модели (Random Forest)
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# Предсказания базовой модели
y_pred = model.predict(X_test)
```

```
# Оценка демографического паритета без корректировки
def evaluate_fairness(y_true, y_pred, sensitive_feature):
    metric_frame = MetricFrame(metrics={"Accuracy": accuracy_score},
                               y_true=y_true,
                               y_pred=y_pred,
                               sensitive_features=sensitive_feature)
    dp_diff = demographic_parity_difference(y_true, y_pred,
sensitive_features=sensitive_feature)
    return metric_frame.by_group, dp_diff

# Результаты без корректировки
accuracy_by_group_no_correction, dp_diff_no_correction =
evaluate_fairness(y_test, y_pred, s_test)

# Применение Fairlearn (Post-processing)
threshold_optimizer = ThresholdOptimizer(
    estimator=model,
    constraints="demographic_parity",
    prefit=True,
    predict_method='predict_proba'
)
threshold_optimizer.fit(X_train, y_train, sensitive_features=s_train)
y_pred_fairlearn = threshold_optimizer.predict(X_test,
sensitive_features=s_test)

# Результаты с Fairlearn
accuracy_by_group_fairlearn, dp_diff_fairlearn =
evaluate_fairness(y_test, y_pred_fairlearn, s_test)

# Применение AI Fairness 360 (Pre-processing)
dataset_orig = BinaryLabelDataset(df=pd.concat([X_train, y_train],
axis=1),
                                  label_names=['label'],
                                  protected_attribute_names=['age'])
dataset_orig.features[:, dataset_orig.feature_names.index('age')] =
(dataset_orig.features[:, dataset_orig.feature_names.index('age')] >
25).astype(int)

rw = Reweighing(unprivileged_groups=[{'age': 0}],
privileged_groups=[{'age': 1}])
dataset_transf = rw.fit_transform(dataset_orig)

# Обучение модели на преобразованных данных
model_aif360 = RandomForestClassifier(random_state=42)
model_aif360.fit(dataset_transf.features,
dataset_transf.labels.ravel())

# Преобразование тестовых данных
dataset_test = BinaryLabelDataset(df=pd.concat([X_test, y_test],
axis=1),
                                  label_names=['label'],
                                  protected_attribute_names=['age'])
dataset_test.features[:, dataset_test.feature_names.index('age')] =
(dataset_test.features[:, dataset_test.feature_names.index('age')] >
25).astype(int)

y_pred_aif360 = model_aif360.predict(dataset_test.features)

# Результаты с AIF360
accuracy_by_group_aif360, dp_diff_aif360 = evaluate_fairness(y_test,
y_pred_aif360, s_test)

# Визуализация результатов
labels = ['No Correction', 'Fairlearn', 'AIF360']
dp_diff_values = [dp_diff_no_correction, dp_diff_fairlearn,
dp_diff_aif360]

plt.figure(figsize=(10, 6))
plt.bar(labels, dp_diff_values, color=['blue', 'green', 'orange'])
plt.title('Demographic Parity Difference')
plt.ylabel('Difference')
plt.show()

# Вывод точности по группам
print("Accuracy by group (No Correction):\n",
accuracy_by_group_no_correction)
print("Accuracy by group (Fairlearn):\n", accuracy_by_group_fairlearn)
print("Accuracy by group (AIF360):\n", accuracy_by_group_aif360)

# What-If Tool (WIT)
# Подготовка данных для WIT
num_datapoints = 1000  # Ограничим количество точек для WIT
X_test_wit = X_test[:num_datapoints].values.tolist()  # Преобразуем
данные в список списков
y_test_wit = y_test[:num_datapoints].values.tolist()  # Преобразуем
метки в список

# Создание конфигурации для WIT
config_builder = WitConfigBuilder(
    examples=X_test_wit,
    feature_names=X_test.columns.tolist()
).set_model_type('classification') \
 .set_target_feature('label') \
 .set_label_vocab(['Denied', 'Approved']) \
 .set_custom_predict_fn(lambda examples:
model.predict_proba(np.array(examples))[:, 1])

# Запуск WIT
WitWidget(config_builder)
```

```
Requirement already satisfied: fairlearn in
/usr/local/lib/python3.11/dist-packages (0.12.0)
Requirement already satisfied: aif360 in
/usr/local/lib/python3.11/dist-packages (0.6.1)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.11/dist-packages (1.6.1)
Requirement already satisfied: matplotlib in
/usr/local/lib/python3.11/dist-packages (3.10.0)
Requirement already satisfied: pandas in
/usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: protobuf==3.20.3 in
/usr/local/lib/python3.11/dist-packages (3.20.3)
Requirement already satisfied: witwidget in
/usr/local/lib/python3.11/dist-packages (1.8.1)
Requirement already satisfied: numpy>=1.24.4 in
/usr/local/lib/python3.11/dist-packages (from fairlearn) (1.26.4)
Requirement already satisfied: scipy>=1.9.3 in
/usr/local/lib/python3.11/dist-packages (from fairlearn) (1.14.1)
Requirement already satisfied: joblib>=1.2.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.1)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.11/dist-packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.11/dist-packages (from pandas) (2025.1)
Requirement already satisfied: absl-py>=0.4 in
/usr/local/lib/python3.11/dist-packages (from witwidget) (1.4.0)
Requirement already satisfied: google-api-python-client>=1.7.8 in
/usr/local/lib/python3.11/dist-packages (from witwidget) (2.160.0)
Requirement already satisfied: ipywidgets>=7.0.0 in
/usr/local/lib/python3.11/dist-packages (from witwidget) (7.7.1)
Requirement already satisfied: oauth2client>=4.1.3 in
/usr/local/lib/python3.11/dist-packages (from witwidget) (4.1.3)
```

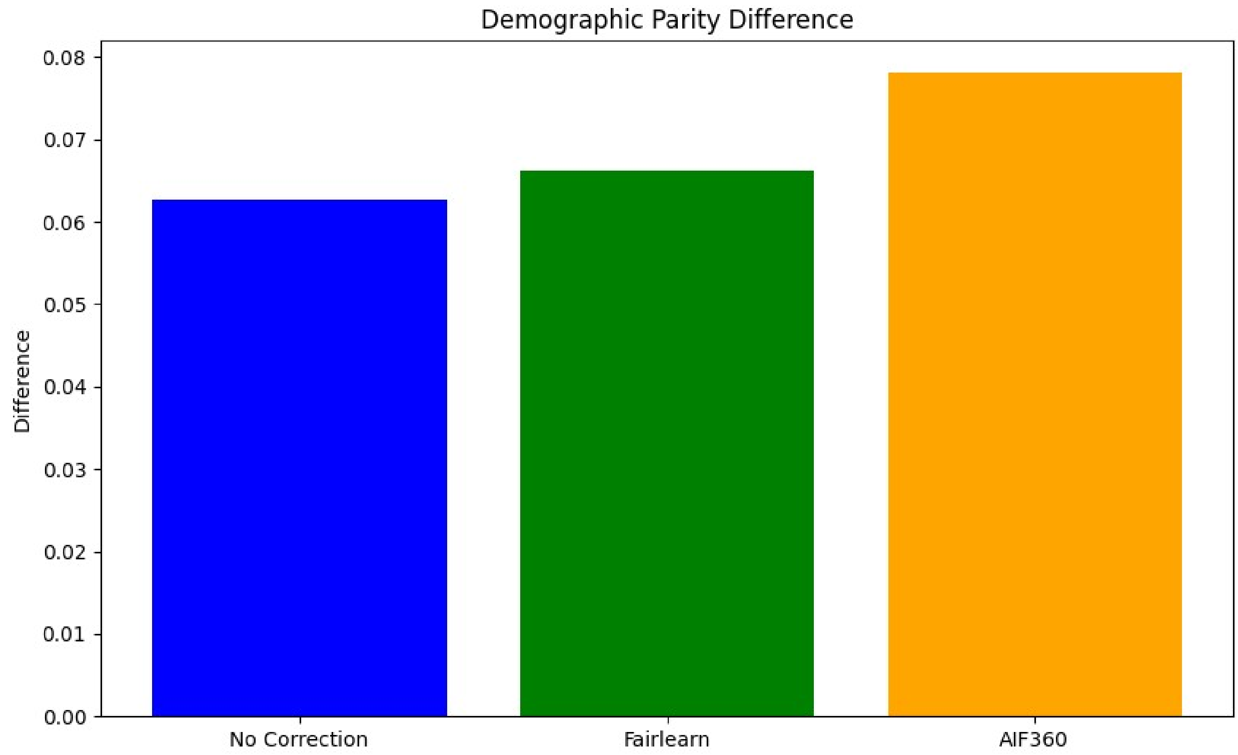
```
Requirement already satisfied: six>=1.12.0 in
/usr/local/lib/python3.11/dist-packages (from witwidget) (1.17.0)
Requirement already satisfied: tensorflow>=1.12.1 in
/usr/local/lib/python3.11/dist-packages (from witwidget) (2.18.0)
Requirement already satisfied: httplib2<1.dev0,>=0.19.0 in
/usr/local/lib/python3.11/dist-packages (from google-api-python-
client>=1.7.8->witwidget) (0.22.0)
Requirement already satisfied: google-auth!=2.24.0,!
=2.25.0,<3.0.0.dev0,>=1.32.0 in /usr/local/lib/python3.11/dist-
packages (from google-api-python-client>=1.7.8->witwidget) (2.38.0)
Requirement already satisfied: google-auth-httplib2<1.0.0,>=0.2.0
in /usr/local/lib/python3.11/dist-packages (from google-api-python-
client>=1.7.8->witwidget) (0.2.0)
Requirement already satisfied: google-api-core!=2.0.*,!=2.1.*,!
=2.2.*,!=2.3.0,<3.0.0.dev0,>=1.31.5 in /usr/local/lib/python3.11/dist-
packages (from google-api-python-client>=1.7.8->witwidget) (2.24.2)
Requirement already satisfied: uritemplate<5,>=3.0.1 in
/usr/local/lib/python3.11/dist-packages (from google-api-python-
client>=1.7.8->witwidget) (4.1.1)
Requirement already satisfied: ipykernel>=4.5.1 in
/usr/local/lib/python3.11/dist-packages (from ipywidgets>=7.0.0-
>witwidget) (6.17.1)
Requirement already satisfied: ipython-genutils~=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from ipywidgets>=7.0.0-
>witwidget) (0.2.0)
Requirement already satisfied: traitlets>=4.3.1 in
/usr/local/lib/python3.11/dist-packages (from ipywidgets>=7.0.0-
>witwidget) (5.7.1)
Requirement already satisfied: widgetsnbextension~=3.6.0 in
/usr/local/lib/python3.11/dist-packages (from ipywidgets>=7.0.0-
>witwidget) (3.6.10)
Requirement already satisfied: ipython>=4.0.0 in
/usr/local/lib/python3.11/dist-packages (from ipywidgets>=7.0.0-
>witwidget) (7.34.0)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in
/usr/local/lib/python3.11/dist-packages (from ipywidgets>=7.0.0-
>witwidget) (3.0.13)
Requirement already satisfied: pyasn1>=0.1.7 in
/usr/local/lib/python3.11/dist-packages (from oauth2client>=4.1.3-
>witwidget) (0.6.1)
Requirement already satisfied: pyasn1-modules>=0.0.5 in
/usr/local/lib/python3.11/dist-packages (from oauth2client>=4.1.3-
>witwidget) (0.4.1)
Requirement already satisfied: rsa>=3.1.4 in
/usr/local/lib/python3.11/dist-packages (from oauth2client>=4.1.3-
>witwidget) (4.9)
Requirement already satisfied: astunparse>=1.6.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (1.6.3)
Requirement already satisfied: flatbuffers>=24.3.25 in
```

```
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (25.2.10)
Requirement already satisfied: gast!=0.5.0,!=0.5.1,!=0.5.2,>=0.2.1
in /usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (0.6.0)
Requirement already satisfied: google-pasta>=0.1.1 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (0.2.0)
Requirement already satisfied: libclang>=13.0.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (18.1.1)
Requirement already satisfied: opt-einsum>=2.3.2 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (3.4.0)
Requirement already satisfied: requests<3,>=2.21.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (2.32.3)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (75.1.0)
Requirement already satisfied: termcolor>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (2.5.0)
Requirement already satisfied: typing-extensions>=3.6.6 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (4.12.2)
Requirement already satisfied: wrapt>=1.11.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (1.17.2)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (1.71.0)
Requirement already satisfied: tensorboard<2.19,>=2.18 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (2.18.0)
Requirement already satisfied: keras>=3.5.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (3.8.0)
Requirement already satisfied: h5py>=3.11.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (3.12.1)
Requirement already satisfied: ml-dtypes<0.5.0,>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (0.4.1)
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in
/usr/local/lib/python3.11/dist-packages (from tensorflow>=1.12.1-
>witwidget) (0.37.1)
Requirement already satisfied: wheel<1.0,>=0.23.0 in
/usr/local/lib/python3.11/dist-packages (from astunparse>=1.6.0-
>tensorflow>=1.12.1->witwidget) (0.45.1)
```

```
Requirement already satisfied: googleapis-common-protos<2.0.0,>=1.56.2
in /usr/local/lib/python3.11/dist-packages (from google-api-core!
=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0.dev0,>=1.31.5->google-api-
python-client>=1.7.8->witwidget) (1.69.1)
Requirement already satisfied: proto-plus<2.0.0,>=1.22.3 in
/usr/local/lib/python3.11/dist-packages (from google-api-core!=2.0.*,!
=2.1.*,!=2.2.*,!=2.3.0,<3.0.0.dev0,>=1.31.5->google-api-python-
client>=1.7.8->witwidget) (1.26.1)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from google-auth!=2.24.0,!
=2.25.0,<3.0.0.dev0,>=1.32.0->google-api-python-client>=1.7.8-
>witwidget) (5.5.2)
Requirement already satisfied: debugpy>=1.0 in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (1.8.0)
Requirement already satisfied: jupyter-client>=6.1.12 in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (6.1.12)
Requirement already satisfied: matplotlib-inline>=0.1 in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (0.1.7)
Requirement already satisfied: nest-asyncio in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (1.6.0)
Requirement already satisfied: psutil in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (5.9.5)
Requirement already satisfied: pyzmq>=17 in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (24.0.1)
Requirement already satisfied: tornado>=6.1 in
/usr/local/lib/python3.11/dist-packages (from ipykernel>=4.5.1-
>ipywidgets>=7.0.0->witwidget) (6.4.2)
Requirement already satisfied: jedi>=0.16 in
/usr/local/lib/python3.11/dist-packages (from ipython>=4.0.0-
>ipywidgets>=7.0.0->witwidget) (0.19.2)
Requirement already satisfied: decorator in
/usr/local/lib/python3.11/dist-packages (from ipython>=4.0.0-
>ipywidgets>=7.0.0->witwidget) (4.4.2)
Requirement already satisfied: pickleshare in
/usr/local/lib/python3.11/dist-packages (from ipython>=4.0.0-
>ipywidgets>=7.0.0->witwidget) (0.7.5)
Requirement already satisfied: prompt-toolkit!=3.0.0,!
=3.0.1,<3.1.0,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from
ipython>=4.0.0->ipywidgets>=7.0.0->witwidget) (3.0.50)
Requirement already satisfied: pygments in
/usr/local/lib/python3.11/dist-packages (from ipython>=4.0.0-
>ipywidgets>=7.0.0->witwidget) (2.18.0)
Requirement already satisfied: backcall in
/usr/local/lib/python3.11/dist-packages (from ipython>=4.0.0-
>ipywidgets>=7.0.0->witwidget) (0.2.0)
Requirement already satisfied: pexpect>4.3 in
/usr/local/lib/python3.11/dist-packages (from ipython>=4.0.0-
>ipywidgets>=7.0.0->witwidget) (4.9.0)
Requirement already satisfied: rich in /usr/local/lib/python3.11/dist-
packages (from keras>=3.5.0->tensorflow>=1.12.1->witwidget) (13.9.4)
Requirement already satisfied: namex in
/usr/local/lib/python3.11/dist-packages (from keras>=3.5.0-
>tensorflow>=1.12.1->witwidget) (0.0.8)
Requirement already satisfied: optree in
/usr/local/lib/python3.11/dist-packages (from keras>=3.5.0-
>tensorflow>=1.12.1->witwidget) (0.14.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2.21.0-
>tensorflow>=1.12.1->witwidget) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2.21.0-
>tensorflow>=1.12.1->witwidget) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2.21.0-
>tensorflow>=1.12.1->witwidget) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2.21.0-
>tensorflow>=1.12.1->witwidget) (2025.1.31)
Requirement already satisfied: markdown>=2.6.8 in
/usr/local/lib/python3.11/dist-packages (from tensorboard<2.19,>=2.18-
>tensorflow>=1.12.1->witwidget) (3.7)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0
in /usr/local/lib/python3.11/dist-packages (from
tensorboard<2.19,>=2.18->tensorflow>=1.12.1->witwidget) (0.7.2)
Requirement already satisfied: werkzeug>=1.0.1 in
/usr/local/lib/python3.11/dist-packages (from tensorboard<2.19,>=2.18-
>tensorflow>=1.12.1->witwidget) (3.1.3)
Requirement already satisfied: notebook>=4.4.1 in
/usr/local/lib/python3.11/dist-packages (from
widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (6.5.5)
Requirement already satisfied: parso<0.9.0,>=0.8.4 in
/usr/local/lib/python3.11/dist-packages (from jedi>=0.16-
>ipython>=4.0.0->ipywidgets>=7.0.0->witwidget) (0.8.4)
Requirement already satisfied: jupyter-core>=4.6.0 in
/usr/local/lib/python3.11/dist-packages (from jupyter-client>=6.1.12-
>ipykernel>=4.5.1->ipywidgets>=7.0.0->witwidget) (5.7.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (3.1.6)
Requirement already satisfied: argon2-cffi in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (23.1.0)
Requirement already satisfied: nbformat in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (5.10.4)
Requirement already satisfied: nbconvert>=5 in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (7.16.6)
Requirement already satisfied: Send2Trash>=1.8.0 in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (1.8.3)
Requirement already satisfied: terminado>=0.8.3 in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (0.18.1)
Requirement already satisfied: prometheus-client in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (0.21.1)
Requirement already satisfied: nbclassic>=0.4.7 in
/usr/local/lib/python3.11/dist-packages (from notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (1.2.0)
Requirement already satisfied: ptyprocess>=0.5 in
/usr/local/lib/python3.11/dist-packages (from pexpect>4.3-
>ipython>=4.0.0->ipywidgets>=7.0.0->witwidget) (0.7.0)
Requirement already satisfied: wcwidth in
/usr/local/lib/python3.11/dist-packages (from prompt-toolkit!=3.0.0,!
=3.0.1,<3.1.0,>=2.0.0->ipython>=4.0.0->ipywidgets>=7.0.0->witwidget)
(0.2.13)
Requirement already satisfied: MarkupSafe>=2.1.1 in
/usr/local/lib/python3.11/dist-packages (from werkzeug>=1.0.1-
>tensorboard<2.19,>=2.18->tensorflow>=1.12.1->witwidget) (3.0.2)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.11/dist-packages (from rich->keras>=3.5.0-
>tensorflow>=1.12.1->witwidget) (3.0.0)
Requirement already satisfied: platformdirs>=2.5 in
/usr/local/lib/python3.11/dist-packages (from jupyter-core>=4.6.0-
>jupyter-client>=6.1.12->ipykernel>=4.5.1->ipywidgets>=7.0.0-
>witwidget) (4.3.6)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0-
>rich->keras>=3.5.0->tensorflow>=1.12.1->witwidget) (0.1.2)
Requirement already satisfied: notebook-shim>=0.2.3 in
/usr/local/lib/python3.11/dist-packages (from nbclassic>=0.4.7-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (0.2.4)
Requirement already satisfied: beautifulsoup4 in
/usr/local/lib/python3.11/dist-packages (from nbconvert>=5-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (4.13.3)
Requirement already satisfied: bleach!=5.0.0 in
/usr/local/lib/python3.11/dist-packages (from bleach[css]!=5.0.0-
>nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (6.2.0)
Requirement already satisfied: defusedxml in
/usr/local/lib/python3.11/dist-packages (from nbconvert>=5-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (0.7.1)
Requirement already satisfied: jupyterlab-pygments in
/usr/local/lib/python3.11/dist-packages (from nbconvert>=5-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (0.3.0)
Requirement already satisfied: mistune<4,>=2.0.3 in
/usr/local/lib/python3.11/dist-packages (from nbconvert>=5-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (3.1.2)
Requirement already satisfied: nbclient>=0.5.0 in
/usr/local/lib/python3.11/dist-packages (from nbconvert>=5-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (0.10.2)
Requirement already satisfied: pandocfilters>=1.4.1 in
/usr/local/lib/python3.11/dist-packages (from nbconvert>=5-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (1.5.1)
Requirement already satisfied: fastjsonschema>=2.15 in
/usr/local/lib/python3.11/dist-packages (from nbformat-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (2.21.1)
Requirement already satisfied: jsonschema>=2.6 in
/usr/local/lib/python3.11/dist-packages (from nbformat-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (4.23.0)
Requirement already satisfied: argon2-cffi-bindings in
/usr/local/lib/python3.11/dist-packages (from argon2-cffi-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (21.2.0)
Requirement already satisfied: webencodings in
/usr/local/lib/python3.11/dist-packages (from bleach!=5.0.0-
>bleach[css]!=5.0.0->nbconvert>=5->notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (0.5.1)
Requirement already satisfied: tinycss2<1.5,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from bleach[css]!=5.0.0-
>nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (1.4.0)
Requirement already satisfied: attrs>=22.2.0 in
/usr/local/lib/python3.11/dist-packages (from jsonschema>=2.6-
>nbformat->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (25.1.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.11/dist-packages (from jsonschema>=2.6-
>nbformat->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.11/dist-packages (from jsonschema>=2.6-
>nbformat->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (0.36.2)
```

```
Requirement already satisfied: rpds-py>=0.7.1 in
/usr/local/lib/python3.11/dist-packages (from jsonschema>=2.6-
>nbformat->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (0.23.1)
Requirement already satisfied: jupyter-server<3,>=1.8 in
/usr/local/lib/python3.11/dist-packages (from notebook-shim>=0.2.3-
>nbclassic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (1.24.0)
Requirement already satisfied: cffi>=1.0.1 in
/usr/local/lib/python3.11/dist-packages (from argon2-cffi-bindings-
>argon2-cffi->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (1.17.1)
Requirement already satisfied: soupsieve>1.2 in
/usr/local/lib/python3.11/dist-packages (from beautifulsoup4-
>nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0-
>ipywidgets>=7.0.0->witwidget) (2.6)
Requirement already satisfied: pycparser in
/usr/local/lib/python3.11/dist-packages (from cffi>=1.0.1->argon2-
cffi-bindings->argon2-cffi->notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (2.22)
Requirement already satisfied: anyio<4,>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from jupyter-server<3,>=1.8-
>notebook-shim>=0.2.3->nbclassic>=0.4.7->notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (3.7.1)
Requirement already satisfied: websocket-client in
/usr/local/lib/python3.11/dist-packages (from jupyter-server<3,>=1.8-
>notebook-shim>=0.2.3->nbclassic>=0.4.7->notebook>=4.4.1-
>widgetsnbextension~=3.6.0->ipywidgets>=7.0.0->witwidget) (1.8.0)
Requirement already satisfied: sniffio>=1.1 in
/usr/local/lib/python3.11/dist-packages (from anyio<4,>=3.1.0-
>jupyter-server<3,>=1.8->notebook-shim>=0.2.3->nbclassic>=0.4.7-
>notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets>=7.0.0-
>witwidget) (1.3.1)
```

```
/usr/local/lib/python3.11/dist-packages/aif360/algorithms/
preprocessing/reweighing.py:66: RuntimeWarning: invalid value
encountered in scalar divide
  self.w_p_fav = n_fav*n_p / (n*n_p_fav)
/usr/local/lib/python3.11/dist-packages/aif360/algorithms/preprocessin
g/reweighing.py:67: RuntimeWarning: invalid value encountered in
scalar divide
  self.w_p_unfav = n_unfav*n_p / (n*n_p_unfav)
/usr/local/lib/python3.11/dist-packages/aif360/algorithms/preprocessin
g/reweighing.py:68: RuntimeWarning: invalid value encountered in
scalar divide
  self.w_up_fav = n_fav*n_up / (n*n_up_fav)
/usr/local/lib/python3.11/dist-packages/aif360/algorithms/preprocessin
g/reweighing.py:69: RuntimeWarning: invalid value encountered in
scalar divide
  self.w_up_unfav = n_unfav*n_up / (n*n_up_unfav)
```



```
Accuracy by group (No Correction):
      Accuracy
age
0    0.741935
1    0.777311
Accuracy by group (Fairlearn):
      Accuracy
age
0    0.693548
1    0.773109
Accuracy by group (AIF360):
      Accuracy
age
0    0.741935
1    0.768908

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<witwidget.notebook.colab.wit.WitWidget at 0x7afb4e1bc090>

/usr/local/lib/python3.11/dist-packages/sklearn/utils/
validation.py:2739: UserWarning: X does not have valid feature names,
but RandomForestClassifier was fitted with feature names
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:27
```

```
39: UserWarning: X does not have valid feature names, but
RandomForestClassifier was fitted with feature names
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:27
39: UserWarning: X does not have valid feature names, but
RandomForestClassifier was fitted with feature names
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:27
39: UserWarning: X does not have valid feature names, but
RandomForestClassifier was fitted with feature names
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:27
39: UserWarning: X does not have valid feature names, but
RandomForestClassifier was fitted with feature names
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:27
39: UserWarning: X does not have valid feature names, but
RandomForestClassifier was fitted with feature names
  warnings.warn(

# Установка библиотеки
!pip install fairlearn

# Импорт библиотек
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from fairlearn.metrics import demographic_parity_difference,
MetricFrame
from fairlearn.postprocessing import ThresholdOptimizer

# Загрузка данных
url =
"https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/
german/german.data"
columns = ['status', 'duration', 'credit_history', 'purpose',
'amount', 'savings', 'employment',
           'installment_rate', 'personal_status', 'other_debtors',
'residence_since', 'property',
           'age', 'other_installments', 'housing', 'credits', 'job',
'people_liable', 'telephone',
           'foreign_worker', 'label']
data = pd.read_csv(url, names=columns, delimiter=' ')
data['label'] = data['label'].apply(lambda x: 1 if x == 1 else 0)

# Разделение данных
X = pd.get_dummies(data.drop('label', axis=1))
y = data['label']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)

# Чувствительный признак (возраст > 25 лет)
s_train = (X_train['age'] > 25).astype(int)
s_test = (X_test['age'] > 25).astype(int)

# Обучение базовой модели
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# Предсказания базовой модели
y_pred = model.predict(X_test)

# Оценка демографического паритета
dp_diff = demographic_parity_difference(y_test, y_pred,
sensitive_features=s_test)
print(f"Demographic Parity Difference (before correction): {dp_diff}")

# Применение ThresholdOptimizer
threshold_optimizer = ThresholdOptimizer(
    estimator=model,
    constraints="demographic_parity",
    prefit=True,
    predict_method='predict_proba'
)
threshold_optimizer.fit(X_train, y_train, sensitive_features=s_train)
y_pred_fairlearn = threshold_optimizer.predict(X_test,
sensitive_features=s_test)

# Оценка демографического паритета после корректировки
dp_diff_fairlearn = demographic_parity_difference(y_test,
y_pred_fairlearn, sensitive_features=s_test)
print(f"Demographic Parity Difference (after correction):
{dp_diff_fairlearn}")
```

```
Requirement already satisfied: fairlearn in
/usr/local/lib/python3.11/dist-packages (0.12.0)
Requirement already satisfied: numpy>=1.24.4 in
/usr/local/lib/python3.11/dist-packages (from fairlearn) (1.26.4)
Requirement already satisfied: pandas>=2.0.3 in
/usr/local/lib/python3.11/dist-packages (from fairlearn) (2.2.2)
Requirement already satisfied: scikit-learn>=1.2.1 in
/usr/local/lib/python3.11/dist-packages (from fairlearn) (1.6.1)
Requirement already satisfied: scipy>=1.9.3 in
/usr/local/lib/python3.11/dist-packages (from fairlearn) (1.14.1)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas>=2.0.3-
>fairlearn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
```

```
/usr/local/lib/python3.11/dist-packages (from pandas>=2.0.3-
>fairlearn) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.11/dist-packages (from pandas>=2.0.3-
>fairlearn) (2025.1)
Requirement already satisfied: joblib>=1.2.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.2.1-
>fairlearn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.2.1-
>fairlearn) (3.5.0)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2-
>pandas>=2.0.3->fairlearn) (1.17.0)
Demographic Parity Difference (before correction): 0.06261859582542695
Demographic Parity Difference (after correction):
0.0016264570344266538
```

Demographic Parity Difference — это метрика, которая измеряет разницу в вероятностях положительного исхода между различными группами (например, по возрасту, полу или другим чувствительным признакам).

. Идеальное значение : Если модель полностью справедлива, то DPD=0. Это означает, что вероятность положительного исхода одинакова для всех групп.