



CDW Documentation

Responsible AI Test

Responsible AI Test

Purpose

Evaluate the Responsible AI dashboard and see what it does.

Test Process

Here's a structured list of **Responsible AI Dashboard Deployment Steps** using the corrected scripts. Each step includes:

- **Step Number & Action**
- **Purpose**
- **Expected Result**

Step 1: Install Required Packages

```
pip install --upgrade raiutils raiwidgets responsibleai ipywidgets
```

Purpose:

Install the Python packages required to run Responsible AI analysis and render the dashboard.

Expected Result:

Packages are installed without errors; dashboard widgets can render in the notebook (after kernel restart).

Step 2: Load and Preprocess the Dataset

```
from sklearn.datasets import fetch_openml
import pandas as pd

data = fetch_openml(name='adult', version=2, as_frame=True)
df = data.frame.dropna()
```

Purpose:

Load a well-known classification dataset (Adult Census Income) and remove any missing values to avoid downstream errors.

Expected Result:

A clean DataFrame with no null values is loaded.

□ Step 3: Split Dataset into Train and Test Sets

```
from sklearn.model_selection import train_test_split

target_column = 'class'
X = df.drop(columns=[target_column])
y = df[target_column]

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

□ Purpose:

Separate features and target, then split into training/testing sets for model training and evaluation.

□ Expected Result:

X_train, X_test, y_train, y_test variables created and stratified properly.

□ Step 4: Define Preprocessing and Train a Model

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier

categorical_cols = X_train.select_dtypes(include=['object',
'category']).columns.tolist()
numerical_cols = X_train.select_dtypes(include=['int64',
'float64']).columns.tolist()

preprocessor = ColumnTransformer([
    ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_cols),
    ('num', StandardScaler(), numerical_cols)
])

clf = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(n_estimators=10, random_state=42))
])

clf.fit(X_train, y_train)
```

□ Purpose:

Build a model pipeline that encodes categorical features, scales numeric ones, and trains a classifier.

□ Expected Result:

Pipeline is trained successfully on the training data without conversion errors.

□ Step 5: Prepare Data for RAIInsights

```
# Ensure target column is a supported type
y_train_clean = y_train.astype(str)
y_test_clean = y_test.astype(str)

train_data = X_train.copy()
train_data[target_column] = y_train_clean

test_data = X_test.copy()
test_data[target_column] = y_test_clean
```

□ Purpose:

Re-attach the target column (as string) to the feature DataFrames — required for RAIInsights.

□ Expected Result:

train_data and test_data DataFrames contain all required columns including the target.

□ Step 6: Initialize the Responsible AI Insights Object

```
from responsibleai import RAIInsights, FeatureMetadata

feature_metadata = FeatureMetadata(categorical_features=categorical_cols)

rai_insights = RAIInsights(
    model=clf,
    train=train_data,
    test=test_data,
    target_column=target_column,
    task_type="classification",
    feature_metadata=feature_metadata
)
```

□ Purpose:

Create a RAIInsights object that acts as the core engine for the Responsible AI dashboard.

□ Expected Result:

RAIInsights object is initialized successfully and ready for configuration.

□ Step 7: Add Responsible AI Analysis Tools

```
rai_insights.explainer.add()
rai_insights.error_analysis.add()
```

```
rai_insights.counterfactual.add(total_CFs=5, desired_class='opposite')
rai_insights.causal.add(treatment_features=categorical_cols)
```

□ Purpose:

Attach various tools (explanation, error analysis, counterfactuals, causal inference) to the insights engine.

□ Expected Result:

No errors thrown; tools are queued for computation.

□ Step 8: Compute Insights

```
rai_insights.compute()
```

□ Purpose:

Run analysis for all selected tools. This step may take a minute or more.

□ Expected Result:

Tool outputs are generated for the first 5,000 rows of the test set.

□ Step 9: Launch the Responsible AI Dashboard

```
from raiwidgets import ResponsibleAIDashboard
ResponsibleAIDashboard(rai_insights)
```

□ Purpose:

Open an interactive dashboard to explore insights such as feature importance, what-if analysis, and error breakdowns.

□ Expected Result:

A dashboard is displayed inside the notebook. Interactive plots and controls are available for analysis.

Download

```
NOTE: Due to the way that these URLs are deployed, this step will fail because the notebook sends the wrong headers and this is expected. You have to either pull the notebook local and use it from the terminal or register the dashboard/dataset and review it through the portal.
```

Output

Error analysis

Tree map **Heat map** **Feature list**

The tree visualization uses the mutual information between each feature and the error to best separate error instances from accurate instances hierarchically in the data. This simplifies the process of discovering and highlighting common feature patterns. To find important feature patterns, look for nodes with a stronger red color (i.e., high error rate) and a higher fit line (i.e., high error coverage). To edit the list of features being used in the tree, click on "Feature list". Use the "Select metric" dropdown menu to learn more about your error and success model performance. Please note that this metric selection will impact the way your error tree is generated.

Basic Information
 All data
 All data (0 items)

Instances in global cohort
 Total: 200
 Correct: 135
 Incorrect: 75

Instances in the selected cohort
 Total: 200
 Correct: 135
 Incorrect: 75

Predictor path (Items)

Error coverage
100.00%

Error rate
15.88%

Model overview

Evaluate the performance of your model by exploring the distribution of your predictor values and the values of your model performance metrics. Use the "Feature cohorts" tab to investigate your model's behavior in a comparative analysis of its performance across different groups of newly created cohort subsets. Use the "Feature cohorts" to investigate your model by looking at a comparative analysis of its performance across sensitive/non-sensitive feature subgroups (e.g., performance across different genders, income levels).

Feature cohorts **Feature cohorts**

Metrics:
 Accuracy score, F1 score, Precision score, Recall score, AUC, ROC

Cohort	Sample size	Accuracy score	F1 score	Precision score	Recall score	AUC
All data	200	0.84	0.82	0.83	0.83	0.98

Probability distribution **Metric visualizations** **Confusion matrix**

Top option chart Choose cohorts

Probability → 0.95

Data analysis

Table view **Chart view**

View the dataset in a table format for all features and rows.

Index	True?	Predicted?	age	workless	salary	education	education-num	sex
5	True	True	41	None	2054	Bachelor	11	Dir
10	True	True	34	None	2054	Some college	10	Ma
11	True	True	41	None	1884	HS grad	9	Ma
13	True	True	43	None	1743	Prof school	10	Dir
22	True	True	47	Local gov	1700	Assoc voc	12	Ma
36	True	True	31	None	1641	HS grad	9	Ma
41	True	True	31	None	2023	Assoc voc	11	Dir
52	True	True	33	None	1204	HS grad	9	Dir

Data analysis

Table view **Chart view**

View the dataset in a table format for all features and rows.

Index	True?	Predicted?	age	workless	salary	education	education-num	sex
5	True	True	41	None	2054	Bachelor	11	Dir
10	True	True	34	None	2054	Some college	10	Ma
11	True	True	41	None	1884	HS grad	9	Ma
13	True	True	43	None	1743	Prof school	10	Dir
22	True	True	47	Local gov	1700	Assoc voc	12	Ma
36	True	True	31	None	1641	HS grad	9	Ma
41	True	True	31	None	2023	Assoc voc	11	Dir
52	True	True	33	None	1204	HS grad	9	Dir

Feature importances

Aggregate feature importance **Individual feature importance**

Explore the top-4 important features that impact your overall model predictions (i.e., global explanations). Use the slider to show descending feature importances. All values feature importances are shown side by side and can be toggled off by selecting the value in the legend. Click any of the features in the graph to see a density plot below of how values of the selected feature affect prediction.

Top 4 features by their importance

Sort by cohort
All data

Chart type
Bar

Line

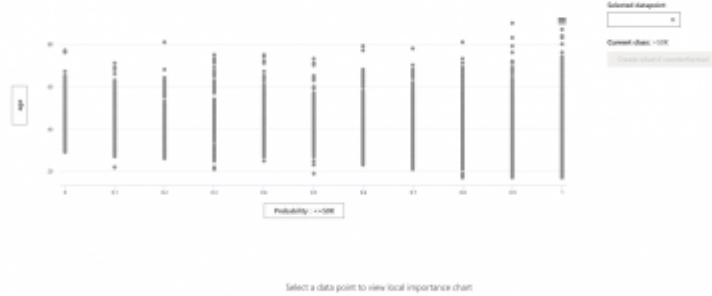
Class importance
Average of absolute

How to read this chart

Show dependence plot for:

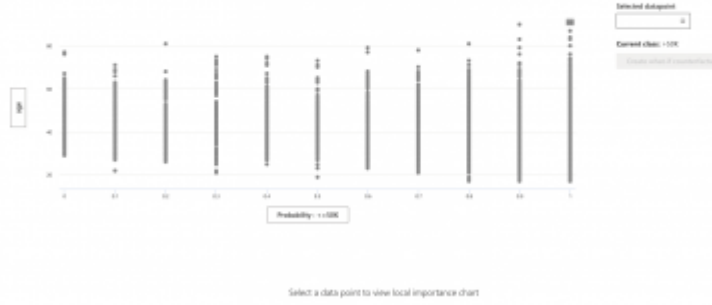
Counterfactuals

What if allow you to perturb features for any input and observe how the model's prediction changes. You can perturb features manually or specify the desired prediction (e.g. class) for a dataset to see a list of closest data points to the original input that would lead to the desired prediction. Also known as prediction counterfactuals, you can use them for exploring the relationships learned by the model, understanding important, necessary features for the model's predictions, or adding edge cases for the model. To start, choose input points from the data table or scatter plot.



Counterfactuals

What if allow you to perturb features for any input and observe how the model's prediction changes. You can perturb features manually or specify the desired prediction (e.g. class) for a dataset to see a list of closest data points to the original input that would lead to the desired prediction. Also known as prediction counterfactuals, you can use them for exploring the relationships learned by the model, understanding important, necessary features for the model's predictions, or adding edge cases for the model. To start, choose input points from the data table or scatter plot.



Causal analysis

The overall causal effects across all data

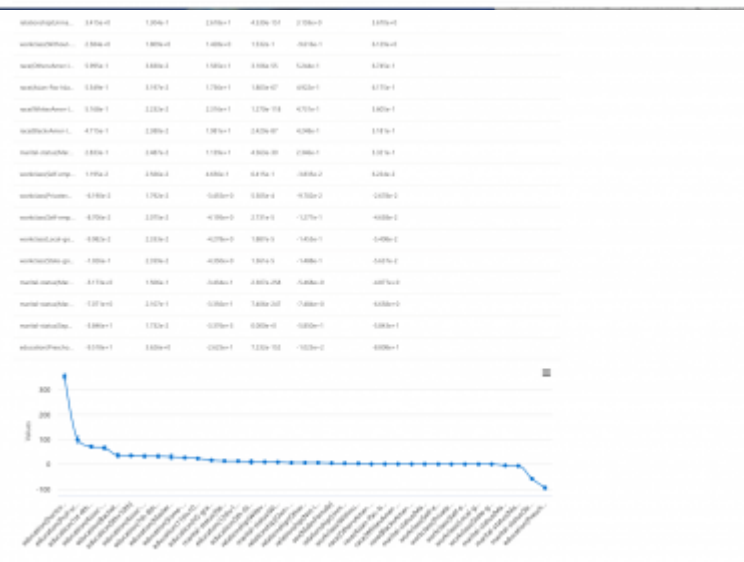
Aggregate causal effects Individual causal effect Treatment policy

Causal analysis assesses "what if" questions about learned relationships. Causal relationships would have changed under different policy options, such as different pricing strategies for a product or an alternative treatment for a patient. Unlike model predictions that identify important conditions or features, these tools help prioritize the most important causal features that directly affect your outcome of interest. These models identify the causal effect of one feature (typically referred to as a "treatment"), holding other confounding features constant. For best results, make sure that the full dataset contains all available features that may confound with the outcome as confounders.

Select aggregate causal effect of each treatment with 95% confidence interval

Why is it important to include confounding features?

Feature	Effect estimate	Standard error	Z score	P-value	Confidence interval	Confidence interval support
education[0,one]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,two]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,three]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,four]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,five]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,six]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,seven]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,eight]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,nine]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,ten]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,eleven]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twelve]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,thirteen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,fourteen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,fifteen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,sixteen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,seventeen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,eighteen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,nineteen]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twenty]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentyone]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentytwo]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentythree]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentyfour]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentyfive]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentysix]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentyseven]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twentyeight]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,twenty-nine]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1
education[0,thirtieth]	0.023e+1	1.88e+0	4.33e+0	0.000e+0	0.023e+1	0.023e+1



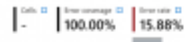
Global cohort: All data (default) Switch cohort New cohort

Error analysis

Tree map [Help map](#)

With the heat map you can focus on specific interactions/feature filters and compare disaggregated error rates. Start with low-impact features to compare.

Cohort: All data



Select metrics: Error rate

Select features: None: Feature 1 Columns: Feature 2

Save as a new cohort

Basic information
All data
All data (2 filters)

Balance in global cohort
Total: 1000
Class 0: 500
Class 1: 500

Balance in the selected cohort
Total: 1000
Class 0: 500
Class 1: 500

Prediction path (filters)

Model overview

Evaluate the performance of your model by exploring the distribution of your predictor values and the values of your model performance metrics. Use the "Global cohort" tab to investigate your model's testing at a comparative analysis of its performance across different pre-built or newly created cohort cohorts. Use the "Feature cohort" to investigate your model by testing at a comparative analysis of its performance across sensitive/non-sensitive feature subcohorts (e.g. performance across different genders, income levels).

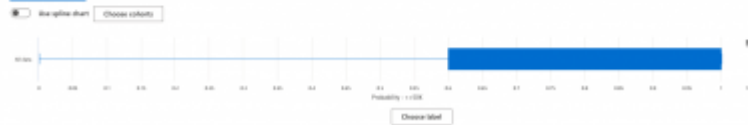
Dataset cohorts: Feature cohorts

Metric
Accuracy score, false positive rate, false negative rate, etc.

Apply the chosen metrics

Cohort	Sample size	Accuracy score	False positive rate	False negative rate	Information
All data	1,000	0.87	0.07	0.03	0.93

Probability distribution **Metric evaluations** Confusion matrix



Data analysis

Table view [Chart view](#)

View the dataset in a table format for all features and rows.

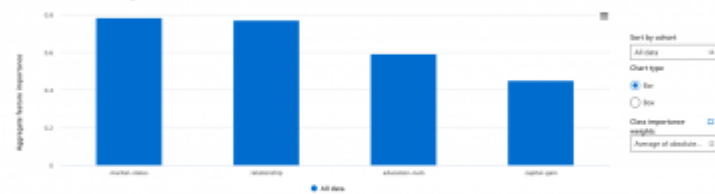
Index	Trust	Predicted	age	workclass	fnbgt	education	education-num	sex
0	100	100	41	Never	20584	Bachelor	16	Male
10	100	100	34	Private	20664	Some-college	16	Male
11	100	100	47	Private	18863	HS grad	9	Male
12	100	100	42	Never	17562	Prof school	12	Male
22	100	100	47	Self-emp	17102	Assoc-voc	12	Male
36	100	100	31	Private	16671	HS grad	9	Male
47	100	100	33	Private	20023	Assoc-voc	11	Male
53	100	100	23	Never	12024	No high	4	Male

Feature importances

Aggregate feature importance **Individual feature importance**

Explore the top-4 important features that impact your overall model/predictions (as a global explanation, use the slider to show descending feature importances). All cohorts' feature importances are shown side-by-side and can be toggled off by selecting the cohort in the legend. Click on any of the features in the graph to see a density plot before or after values of the selected feature affect prediction.

Top-4 features by their importance



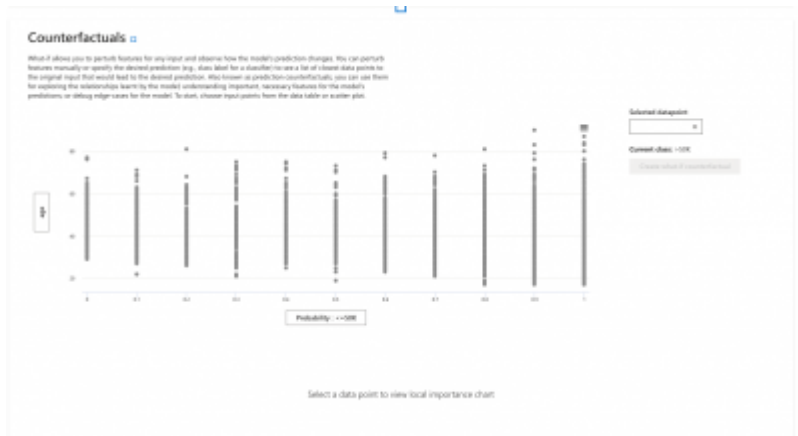
Select a feature to show its dependence plot

View dependence plot for:

Select feature:

Select a depend cohort:

All data



AI Knowledge