Supplementary material
JUICE: JUstIfied Counterfactual Explanations

Anonymous

Anonymous Institution

1 Appendix

1.1 Datasets

Synthetic datasets We report the features of the two synthetic datasets that were generated containing a full mixed-features space.

1. **Disease**: The 7 features of the Disease synthetic dataset are described in Table 1.

Feature	Description	Type	Mutable	Direction	Risk Contribution
Age	Age in years	Continuous	No	-	+0.01 / year
Smokes	Whether smokes or not	Binary	Yes	Any	+0.40
Exercise Minutes	Daily average exercise time (in minutes)	Continuous	Yes	Any	-0.01 / minute
Sleep Hours	Daily average sleep time (in hours)	Continuous	Yes	Any	-0.01/ hour
Weight	Weight category: (1): Underweight, (2): Normal, (3): Overweight, (4): Obese	Ordinal	Yes	Any	(1): +0.05 (2): 0 (3): +0.15 (4): +0.25
Diet	Diet type: (1): Vegan, (2): Vegetarian, (3): Pescetarian, (4): Omnivore	Categorical	Yes	Any	(1): -0.10 (2): -0.05 (3): 0.05 (4): 0.15
Stress	Stress level: (1): Low, (2): Normal, (3): High	Categorical	Yes	Any	(1): -0.10 (2): 0 (3): +0.20

Table 1: Description of the features in the Disease synthetic dataset.

The right column, *Risk contribution*, indicates how much each value of the feature contributes to the risk of being diagnosed with a severe disease. In

the *Age* feature, the yearly 0.01 increase in risk starts from $Age = 40$ and increase linearly until $Age = 100$, meaning that this risk contribution goes from 0% at age 40 to a limit of 60% at age 100 and onwards.
A person who is 42 years old, smokes, exercises 10 minutes daily in average, sleeps 8 hours daily in average, is underweight, omnivore and has a high stress level, has a total risk of developing a severe disease of: $0.02 - 0.1 - 0.08 + 0.05 + 0.15 + 0.20 = 0.24$. The ground truth label for this person is then obtained by drawing either 0 or 1, with probability 0.76 and 0.24, respectively, where 1 means that the person will be diagnosed with a severe disease.

2. **Athlete**: The 6 features of the Athlete dataset are described in Table 2.

Feature	**Description**	**Type**	**Mutable**	**Direction**	**Chances Contribution**
Age	Age in years	Continuous	No	-	-0.01 / year
Gender	Male or Female	Binary	No	-	0
Sport	Sport: (1): Football, (2): Running, (3): Swimming, (4): Shooting	Categorical	Yes	Any	(1): 0 (2): +0.05 (3): +0.10 (4): +0.20
Weekly Training	Weekly training sessions: (1): 3, (2): 4, (3): 5, (4): 6	Ordinal	Yes	Any	(1): -0.15 (2): -0.10 (3): +0.15 (4): +0.30
Sleep Hours	Daily average sleep time (in hours)	Continuous	Yes	Any	+0.01/ hour
Diet	Diet type: (1): Vegan, (2): Vegetarian, (3): Pescetarian, (4): Omnivore	Categorical	Yes	Any	(1): +0.20 (2): +0.15 (3): -0.05 (4): -0.10

Table 2: Description of the features in the Athlete synthetic dataset.

The right column, *Chances contribution*, indicates how much each value of the feature contributes to the chances of winning an olympic medal. In the *Age* feature, the 0.01 yearly decrease in the chance starts from $Age = 25$ and decreases linearly until $Age = 50$. This chance contribution stays at 30% from ages 25 and below, and decreases linearly to a limit of 5% at age 50 and onwards.
An athlete who is 32 years old, female, specializes in air rifle, exercises 5 times a week, sleeps 8 hours daily in average, and is vegetarian, has a total chance of winning an olympic medal of: $0.23 + 0 + 0.20 + 0.15 + 0.08 + 0.15 = 0.80$. The ground truth label for this person is then obtained by drawing either

0 or 1, with probability 0.20 and 0.80, respectively, where 1 means that the person will win an olympic medal.
We make these two synthetic datasets publicly available (see files *disease.csv* and *athlete.csv*).

Publicly available datasets We make a brief description of the publicly available datasets and their respective preprocessing steps.

1. **Ionosphere**: UCI ML Repository dataset available at the website[1] for the prediction of ionospheric condition prediction. A RF model is implemented to obtain 8 features out of the 34 available continuous features according to Mean Decrease in Impurity (MDI) measure. Fig. 1 shows the MDI for all 34 features. The features corresponding to the highest MDI are the most important. In this case, features 0, 2, 4, 5, 6, 7, 26 and 30 are selected.

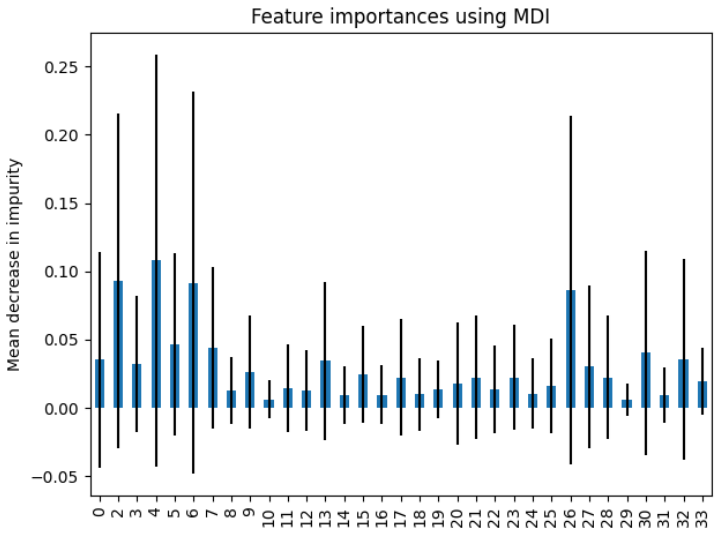

Fig. 1: MDI for all Ionosphere dataset features.

2. **Compass**: Propublica dataset for recidivism prediction, available at the Propublica website[2]. The dataset used is the compass-scores-two-years.csv. The dataset is processed to contain only 5 features, 3 binary, namely Sex $\in$ {Male, Female}, Race $\in$ {African-American, Caucasian}, ChargeDegree $\in$ {Misdemeanor, Felony}, and 2 continuous, namely PriorsCount and Age. The target variable is a new criminal sentence in the next two years.

[1] https://archive.ics.uci.edu/ml/datasets/ionosphere
[2] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

3. **Credit**: UCI ML Repository dataset available at the website[3] for the prediction of default status of credit card holders. The dataset initially contains 24 features, which are preprocessed to obtain a final set of 14 features, 3 binary, namely isMale $\in$ {0,1}, isMarried $\in$ {0,1}, HasHistoryOfOverduePayments $\in$ {0,1}, 7 continuous, namely MaxBillAmountOverLast6Months, MaxPaymentAmountOverLast6Months, MonthsWithZeroBalanceOverLast6Months, MonthsWithLowSpendingOverLast6Months, MonthsWithHighSpendingOverLast6Months, MostRecentBillAmount, MostRecentPaymentAmount and 4 ordinal, namely TotalOverdueCounts $\in$ {0, 1, 2, 3}, TotalMonthsOverdue $\in$ {0, 1, 2, ... 33, 34, 35, 36}, AgeGroup $\in$ {< 25, $\geq 25 \wedge < 40$, $\geq 40 \wedge < 59$, ≥ 60} and EducationLevel $\in$ {High School, University, Graduate, Others}.
4. **Adult**: UCI ML Repository dataset available at the website[4] for adults income prediction. The dataset initially contains 14 features, which are preprocessed to obtain a final set of 12 features, 2 binary, namely Sex $\in$ {Male, Female}, NativeCountry $\in$ {United-States, Non-United-States}, 4 categorical, namely WorkClass $\in$ {Federal-gov, Local-gov, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay}, MaritalStatus $\in$ {Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed}, Occupation $\in$ {Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving} and Relationship $\in$ {Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife}, 5 continuous, namely Age, EducationNumber, CapitalGain, CapitalLoss and HoursPerWeek and 1 ordinal, namely EducationLevel $\in$ {prim-middle-school, high-school, HS-grad, Some-college, Bachelors, Masters, Doctorate, Assoc-voc, Assoc-acdm, Prof-school}.
5. **German**: UCI ML Repository dataset available at the website[5] for credit risk prediction. The dataset initially contains 20 features, which are preprocessed to obtain a final set of 4 features, 1 binary, namely Sex $\in$ {Male, Female} and 3 continuous, namely Age, Credit and LoanDuration.
6. **Heart**: UCI ML Repository dataset available at the website[6] for heart disease prediction. The dataset initially contains 75 features, which are preprocessed to obtain a final set of 7 features, 1 binary, namely Sex $\in$ {Male, Female}, 2 categorical, namely ChestPain $\in$ {1, 2, 3, 4} and ECG $\in$ {0, 1, 2}, and 4 continuous, namely Age, RestBloddPressure, Chol, BloodSugar.

The preprocessing of the **Compass**, **Credit**, **Adult** and **German** datasets is carried out according to the pipeline presented by Karimi et al. [1].

1.2 Model hyperparameters

Table 3 shows the selected models with their corresponding hyperparameters.

[3] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[4] https://archive.ics.uci.edu/ml/datasets/adult
[5] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[6] https://archive.ics.uci.edu/ml/datasets/heart+disease

Table 3: Black-box models hyperparameters

Model	Dataset	Hyperparameters			
		Max. Depth	Samples leaf	Samples split	Trees
RF	Disease	10	3	10	50
	Compass	5	1	5	50
	Adult	10	1	5	100
	Heart	5	5	5	50
		Activation	Hid. Layers	Nodes	Solver
MLP	German	ReLU	2	100, 10	SGD
	Credit	Tanh	2	50, 1	Adam
	Ionosphere	ReLU	2	100, 10	Adam
	Athlete	ReLU	2	100, 10	Adam

2 r Values evaluated

Assume n features, n_{cont} continuous features, and a set of instances from the dataset a synthetic dataset uniformly sampled in the continuous feature space. For this set of instances, we generate a distance matrix DM among all instance. The 18 r values considered for the $r - ball$ in the continuous feature subspace justification verification process are shown below. r_{14} is the vector of minimum values per row of matrix DM. r_5 was chosen as the best r value.

1. $r_1 = \sqrt{n}/5$
2. $r_2 = \sqrt{n}/10$
3. $r_3 = \sqrt{n}/50$
4. $r_4 = \sqrt{n}/100$
5. $r_5 = \sqrt{n_{cont}}/5$
6. $r_6 = \sqrt{n_{cont}}/10$
7. $r_7 = \sqrt{n_{cont}}/50$
8. $r_8 = \sqrt{n_{cont}}/100$
9. $r_9 = max([minPerRow(DM)])$
10. $r_{10} = r_9/5$
11. $r_{11} = r_9/10$
12. $r_{12} = r_9/50$
13. $r_{13} = r_9/100$
14. $r_{14} = min(DM[0,:])$
15. $r_{15} = r_{14} + (r_{14} + max(DM[0,:]))/5$
16. $r_{16} = r_{14} + (r_{14} + max(DM[0,:]))/10$
17. $r_{17} = r_{14} + (r_{14} + +max(DM[0,:]))/50$
18. $r_{18} = r_{14} + (r_{14} + +max(DM[0,:]))/100$

References

1. Karimi, A., Gilles, B., Borja, B., and Isabel, V. *Model-agnostic counterfactual explanations for consequential decisions*, International Conference on Artificial Intelligence and Statistics, (2020), pp. 895–905.