AUTO-DataGenCARS
Study cases

Last update: 16/06/2021

Index

Extend a dataset and check that both datasets (original and extended) are comparable **4**

Add context to a dataset where there is none **7**

Generate a dataset that behaves the same as another but has no bias **11**

Annexes **16**

In this document, different study cases and a validation / testing experiment will be presented; all of them have been resolved using the *AUTO-DataGenCARS* tool.

1. Extend a dataset and check that both datasets (original and extended) are comparable

A reduced *MovieLens 100k*[1] dataset will be used for this experiment, with a total of 20,000 ratings. The histogram of these ratings is shown in Fig. 1. This graph has been created using the Power BI service.

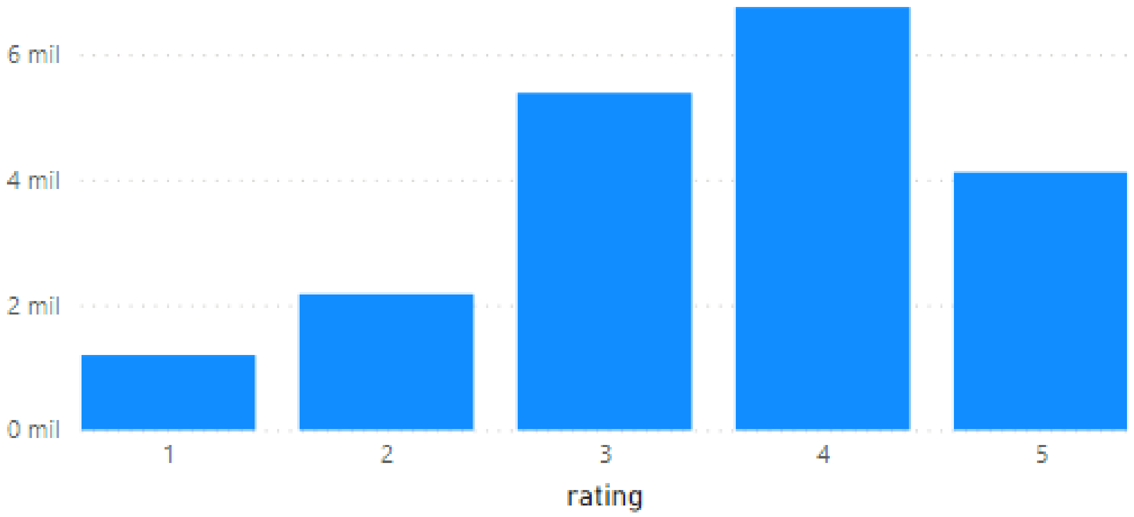


Figure 1. Ratings from the *MovieLens 20k* dataset

The error metric "*MAE*" and the ranking metrics "*Precision*, *Recall* and *F-Measure*" of this dataset will now be calculated using the evaluation frame implemented in the *AUTO-DataGenCARS* tool. To calculate them, *User-Based*[2] and *Item-Based*[3] generic recommenders will be used, in addition to *SVD*[4] (Singular Value Decomposition). Both the user-based and item-based recommenders mentioned are the *Apache Mahout*[5] implementations of *IBCF* (item-based collaborative filtering) and UBCF (user-based collaborative filtering), respectively. These implemented recommenders are user-based and item-based approaches of neighborhood methods, which are the most established and widely used approach to collaborative filtering.

[1] https://www.kaggle.com/prajitdatta/movielens-100k-dataset

[2] GenericUserBasedRecommender (Mahout Map-Reduce 0.13.0 API)

[3] GenericItemBasedRecommender (Mahout Map-Reduce 0.13.0 API)

[4] SVDRecommender (Mahout Map-Reduce 0.13.0 API)

[5] https://mahout.apache.org/

	MAE	Precision	Recall	F-Measure
Generic Item-Based Recommender	1,170	0,002	0,013	0,003
Generic User-Based Recommender	0,934	0,004	0,020	0,007
SVD	1,184	0,007	0,048	0,013

Figure 2. Error and Ranking metrics

A 10-fold cross validation has been used in all these experiments.

This dataset was then expanded from 20,000 to 30,000 total ratings using an "*Increment dataset generation*". To do this, both a user profile file [1] containing the characteristics of this dataset and an item scheme [7] are needed. Using the *AUTO-DataGenCARS* tool, we can easily generate the necessary user profiles by using the "*Generate a dataset's user profile*" option in the "*Workflow menu*" tab and selecting the movie [2] and ratings [5] files in the "*Data input > Preexisting Files*" tab; and write the item scheme according to the data in the items file in the "*Data input > Items*" tab, creating each movie attribute as needed.This scheme will be used as a support file for the expansion, and in this case it will be filled with the data of the columns that make up the original item file (in this case different genres of movies, each attribute taking the value of 1 if a movie is of a certain genre and 0 if it is not).

In short, selecting both "*Generate a dataset's user profile*" and "*Increment dataset generation*" in the workflow's menu, indicating in the "*Generation options*" tab the number of total ratings for out new dataset, selecting the original CSV of the MovieLens dataset ([2], [3], [5]) in the "*Data input>Preexisting files*" tab and filling the scheme in the "*Data input>Items*" tab with different movie genres will allow us to run the generations and expand the dataset.

AUTO-DataGenCARS automatically generates and shows several graphs in the "*Data visualization*" tab after the dataset generation is completed. One of these graphs is the distribution of the ratings generated, which can be seen in Fig. 3.

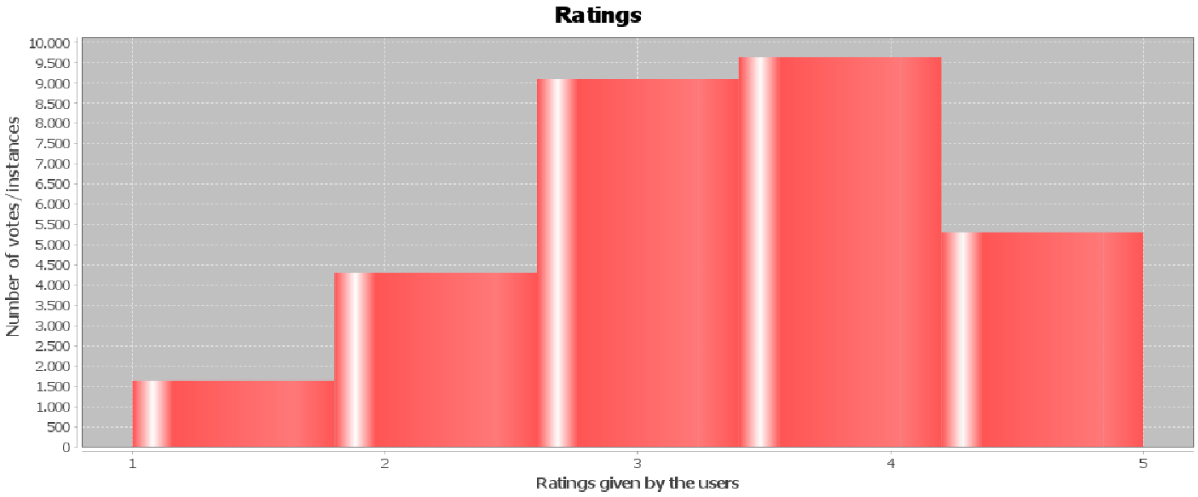


Figure 3. Ratings from the expanded dataset

And the metrics, obtained in the same way as previously explained, are the following:

	MAE	Precision	Recall	F-Measure
Generic Item-Based Recommender	1,620	0,001	0,004	0,001
Generic User-Based Recommender	1,105	0,004	0,012	0,006
SVD	1,024	0,007	0,033	0,011

Figure 4. Metrics from the incremented dataset

Comparing these metrics (as seen in Fig. 5), we can verify that both datasets, the original and the expanded one, are comparable. The attribute values themselves were also compared between these two datasets (average values,standard deviation, percentage of unique values) using the *AUTO-DataGenCARS* built in graphs, and the results obtained are also similar.

	MAE Original/Expanded	Precision Original/Expanded	Recall Original/Expanded	F-Measure Original/Expanded
Generic User-Based Recommender	0,934 / 1,105	0,004 / 0,004	0,020 / 0,012	0,007 / 0,006
SVD	1,184 / 1,024	0,007 / 0,007	0,048 / 0,033	0,013 / 0,011

Figure 5. Metrics comparison

2. Add context to a dataset where there is none

For this experiment, the *MovieLens 100k*[6] dataset was used, which has no context. The following context attributes will be added:

- *room_comfort* (very good, good, adequate, bad, very bad)
- *room_occupancy* (high, medium, low)
- *weather* (sunny, windy, rainy, cloudy)
- *room_cleanliness* (high, medium, low)
- *place_in_room* (far for the screen, close to the screen, center, corner of the room)
- *eating_allowed* (only drinks, only food, food and drinks, none)

To do this, the context scheme [8] will first be created using the *AUTO-DataGenCARS* tool, and use this file to create the contexts data file [4]. The scheme can be filled with the necessary context attributes in the "*Data input > Contexts*" tab, and in the "*Generation options*" tab the number of contexts to be generated can also be chosen, in this case we'll be generating 1000 context instances. This newly created file will be used to perform a synthetic generation of a dataset from an original dataset.

Several user profiles will be generated as well, in the "*Data input > Attribute's weight*" tab. In this tab, the newly created contexts' attributes can be marked with the "*Important weight*" box, which will allow different weights to be given to the marked attributes when creating a user profile. In the "*Preexisting files*" tab the file containing the items of the original dataset, in this case the genres of the movies, can be browsed and selected. This action will make these attributes appear in the "*Attribute's weight*" tab, along with the contexts attributes previously mentioned, and check the "*Important weight*" box in the attributes needed for the user profile. In this case, we will mark every context attribute and three different genres.

Next, click on the "*Create user profile*" button to create different user profiles with these attributes. In this case, several profiles have been created, each one of them with more weight on a specific attribute (e.g. A profile where the occupancy of the movie theater is given more importance, or a specific genre of a film) than the rest. These user profiles can be customized as needed for each case.

Using *AUTO-DataGenCARS* this can be done in a single workflow, which would be as shown in Fig. 6. For the workflow to function this way, we will need to select in the "*Workflow menu*" tab both the "*Context file generation*" and the "*Synthetic dataset generation from CSV*" options, and place them in that order. Filling the context scheme previously mentioned, creating the customized user profiles [6] and selecting the users [3], movies [2] and ratings [5] files in the "*Data input > Preexisting Files*" tab will allow us to run the generation.

[6] https://www.kaggle.com/prajitdatta/movielens-100k-dataset

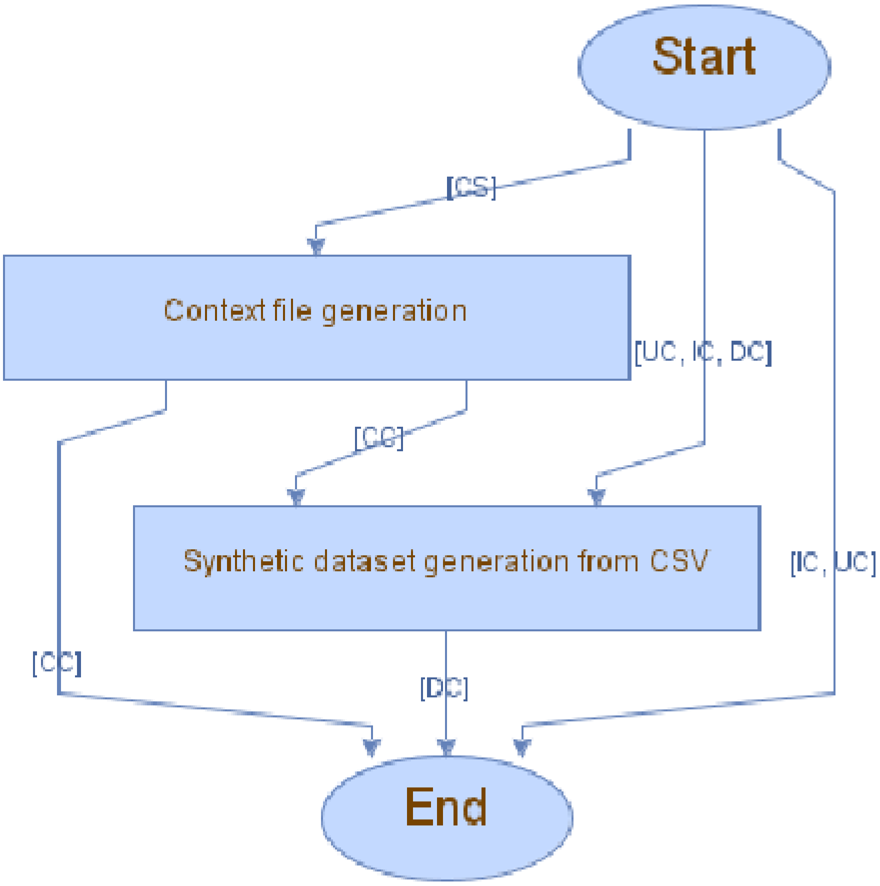


Figure 6. Workflow actions

After running the workflow and generating the data, the ratings' distribution would be as indicated in Fig. 7. This graph, and several others, are created automatically after the dataset generation ends and shown in the "*Data visualization*" tab.

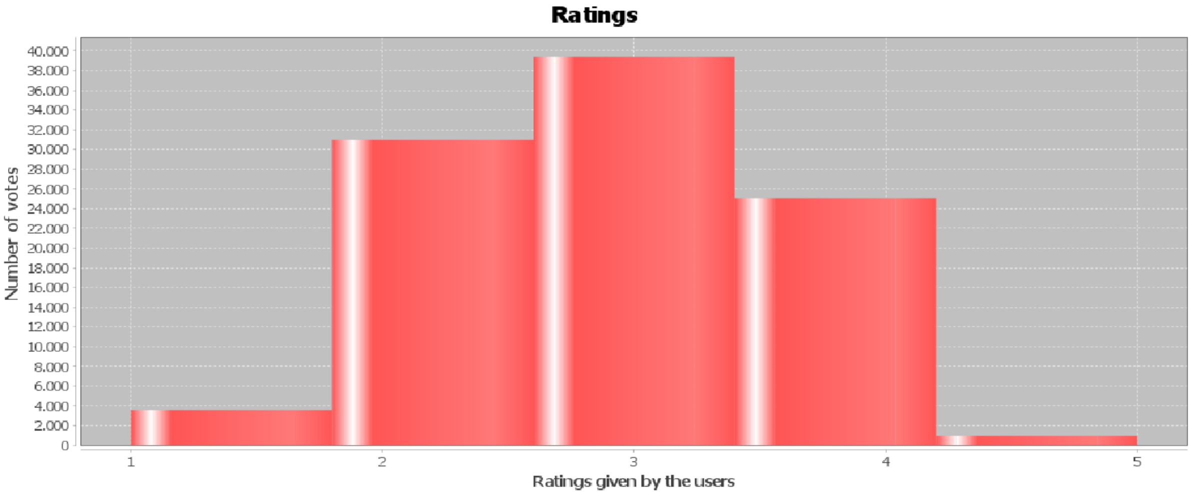


Figure 7. Ratings from the new dataset with context

Next, the impact that this generated context has on the dataset will be evaluated, using the evaluation tool with experimental settings implemented in *AUTO-DataGenCARS*, in the evaluation window. For this, four datasets will be generated (five counting the original), each one with a different number of available contexts values but with the same number of ratings (100,000).

These datasets have been generated with the *AUTO-DataGenCARS* evaluation tool, eliminating in four instances of the original dataset 80%, 60%, 40% and 20% of the available context randomly, respectively. As an example, if the amount of context data in a dataset is 80% this means that 20% of the context variables (randomly selected) do not have a value.

To do this, click on the "*Evaluate this dataset*" button in the "*Data visualization*" tab. In this new menu, click on the *"More experiments"* arrow, the window will expand to be able to carry out more exhaustive experiments with the same options selected at the top of the menu. In this case we are going to perform an evaluation test of the dataset, but only with *200, 400, 600, 800 and 1,000* context instances. By selecting the *x* value "*number of contexts*", specifying that the maximum number of contexts is *1,000* and the increment is done in steps of *200*, this experiment can begin.

This evaluation has been done using a 10-fold cross validation and a SVD recommender.

	MAE	Precision	Recall	F-Measure
100% Context available	0,1762	0,850	0,810	0,793
80% Context available	0,1888	0,819	0,755	0,729
60% Context available	0,2012	0,791	0,704	0,662
40% Context available	0,2135	0,765	0,653	0,589
20% Context available	0,2263	0,735	0,602	0,503

Figure 8. Context evaluation with SVD

Fig. 8. shows that the MAE increases when a higher portion of context is unknown and the F-measure decreases. Fig.9 shows the metrics obtained with a context modeling recommender (Naive bayes), which gives us the same conclusion.

	MAE	Precision	Recall	F-Measure
100% Context available	0,0226	0,984	0,985	0,984
80% Context available	0,0652	0,906	0,897	0,894
60% Context available	0,1081	0,838	0,809	0,798
40% Context available	0,1506	0,775	0,723	0,695
20% Context available	0,1943	0,694	0,634	0,572

Figure 9. Context evaluation with CM

3. Generate a dataset that behaves the same as another but has no bias

For this experiment, the *COMPAS*[7] dataset was used, which has been proven that is biased in favor of white defendants, and against black inmates, when calculating the risks of recidivism[8]. We can see in Fig. 10. how there is much more data on African-American people than the rest of the races, and that in addition a much higher risk of recidivism has been calculated than the rest. Fig. 11. also shows how the likelihood of reoffense calculated is distributed. To carry out this experiment, low recidivism risk values have been substituted for the value of 1, medium for 2, high for 3. The graphs shown in Fig. 10, 11 and 15 have been created using the *Power BI*[9] service.

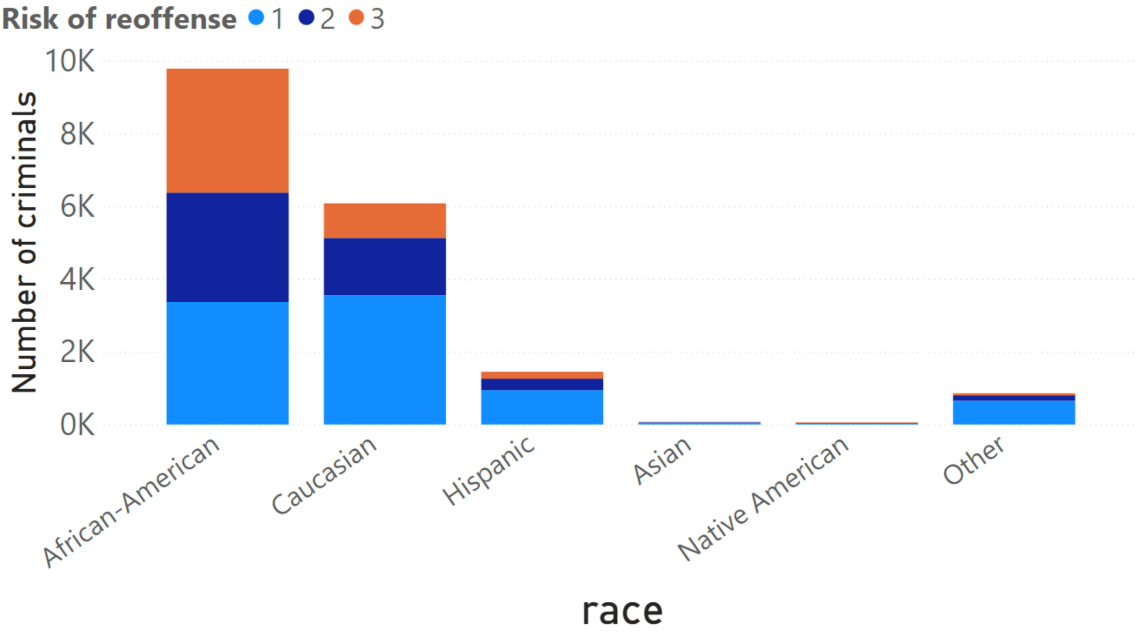


Figure 10. Risks with the unbalanced COMPAS dataset.

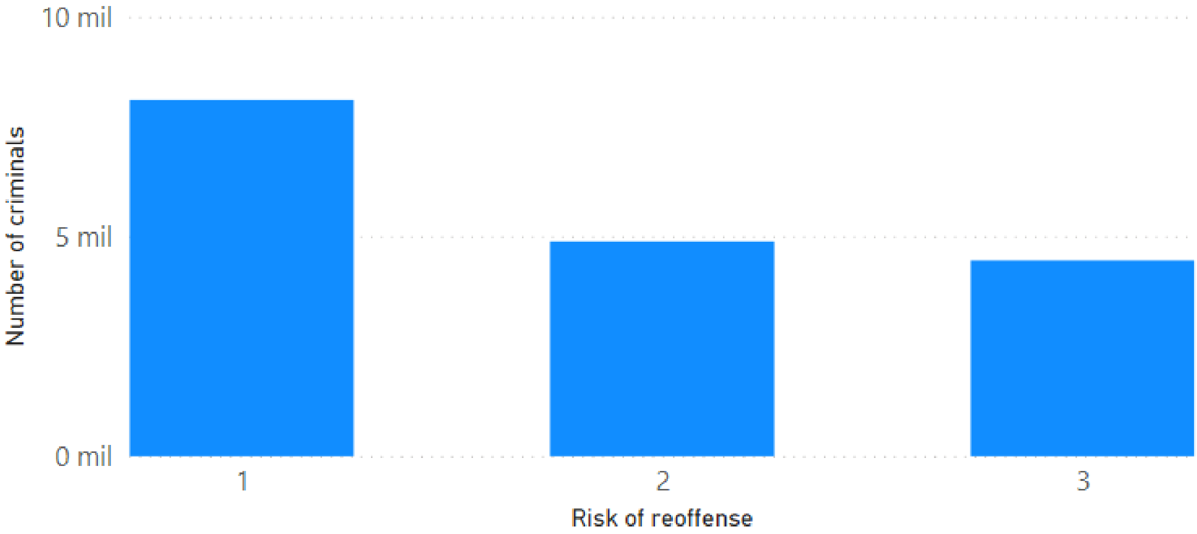
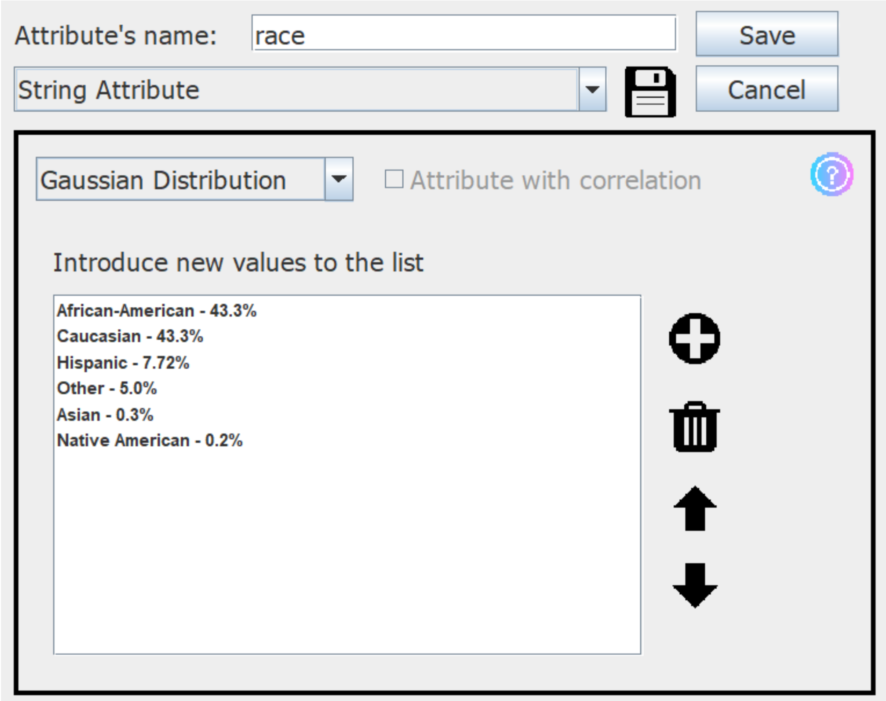


Figure 11. Risks distribution of the unbalanced COMPAS dataset.

[7] https://www.kaggle.com/danofer/compass
[8] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[9] https://powerbi.microsoft.com/

To generate a dataset that behaves the same as the original but without bias, items, which in this case are criminals, will be generated synthetically using the *AUTO-DataGenCARS* tool. These new criminals will be generated with the same proportions and attributes as the original ones, but modifying the percentages of appearance of each race, that is, increasing the number of caucasian prisoners and reducing that of african-americans, as can be seen in Fig. 12. The reoffense risk will be generated with this new data but using the behavior of the original dataset, using the statistics and user profiles of said dataset.



Attribute's name: race

String Attribute

Gaussian Distribution

☐ Attribute with correlation

Introduce new values to the list

African-American - 43.3%
Caucasian - 43.3%
Hispanic - 7.72%
Other - 5.0%
Asian - 0.3%
Native American - 0.2%

Figure 12. Race attribute in the item's scheme

To do this, several steps have been taken. First, the user profiles [1] of the original files of this dataset have been generated, thus storing the behavior of the original *COMPAS* dataset in a file. This action is carried out by the generation option "*Generate a dataset's user profile*" in the first tab of the workflow, using the original dataset files selected in the "*Data input> Preexisting files*" tab. This will create several user profile instances, each one of them representing the importance that a user gives to each of the attributes of the original dataset.

At the same time, a new set of items has been generated (using the attributes created in the "*Data input* > *Items*" tab), which in this case is the information of the prisoners, but changing the percentages of appearance of each race (Fig. 12). In addition to the attribute of the race, all the necessary attributes have been generated so that the new criminals generated are as close to the originals as possible, except for the modified race. To generate as many criminals as the original dataset, you need to indicate the specific number of items to generate, which in this case will be 18,316, in the "*Generation options*" tab.

With the user profiles created and the new prisoners [9] generated, the new risks of recidivism can be generated using the original user files, contexts [10] and "ratings" [5], which in this case are the reoffense risks of the new criminals. All these steps can be performed at the same time in the same workflow if in the "*Workflow Menu*" tab the options

“*Generate a dataset's user profile*”, “*Item generation*” and “*Synthetic dataset generation with CSV*” are selected, in that order.

The workflow would perform the actions indicated in Fig. 13. in that specific order. Fig.13.’s graph can be easily generated in the “*Workflow Menu*” tab by clicking on the “*Current Workflow’s Graph*”, and shows the actions that will be taken by the workflow, as well as the input and output files each action has.

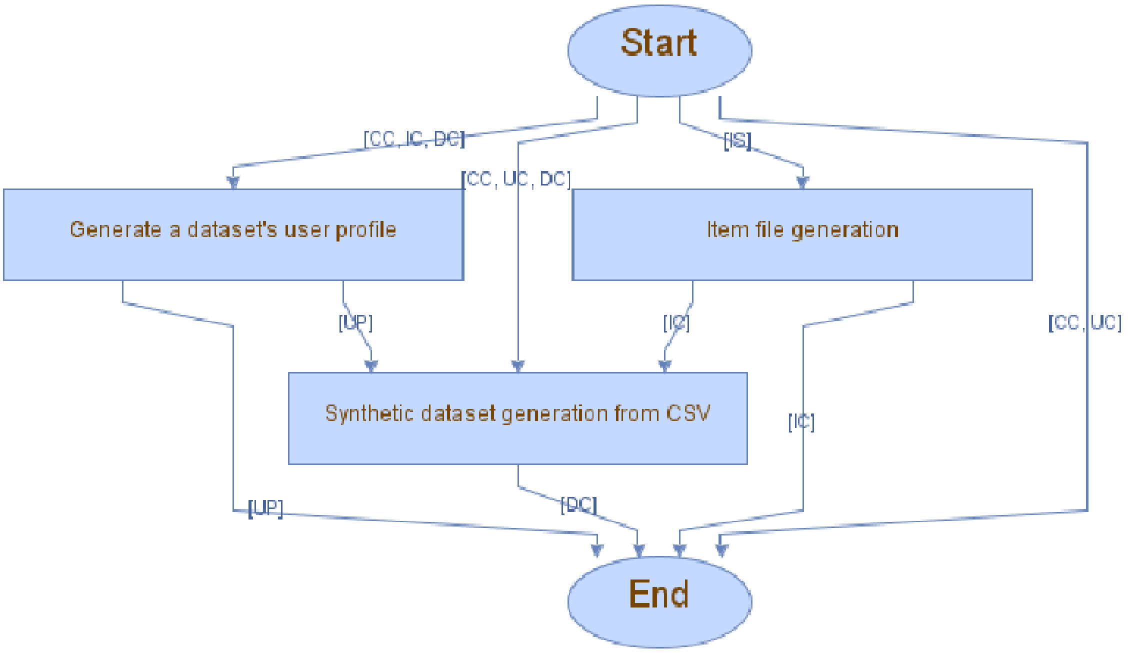


Figure 13. AUTO-DataGenCARS workflow’s actions

When generating the new data, the distribution of the new recidivism risks would be as indicated in Fig. 14. This graph has been created automatically by *AUTO-DataGenCARS* once the dataset has been generated, and it is visible in the “*Data visualization*” tab.

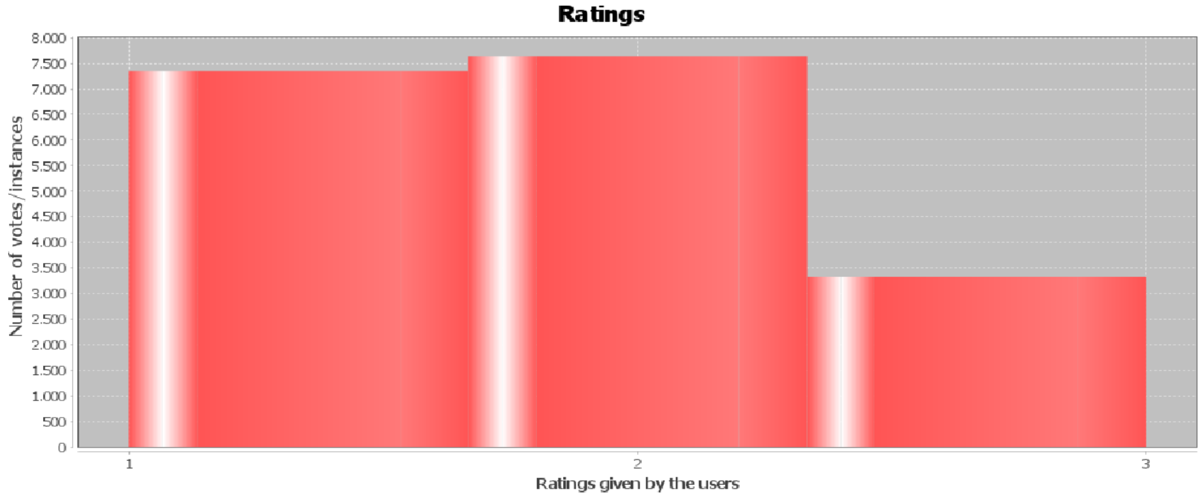


Figure 14. Risks distribution of the balanced COMPAS-based dataset.

Fig. 15 also shows how the number of criminals of both races and their calculated risk of recidivism have been balanced. Comparing both Fig. 10 and 15 it can be seen that the race attribute no longer influences the generation of reoffense risks, so in this new dataset there would no longer be biased in favor of white defendants, and against black inmates.

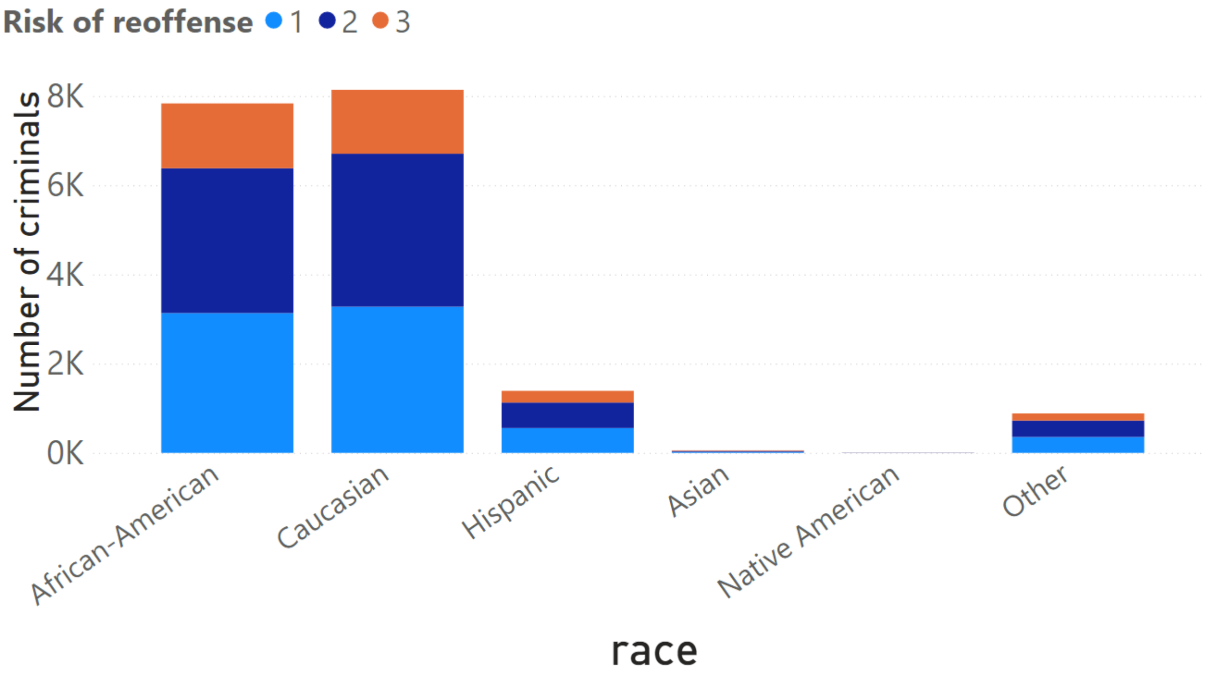


Figure 15. Risks with the balanced COMPAS-based dataset.

Figures 16 and 17 represent the distribution of reoffense risks according to the prior registered crimes of a criminal from the original and the newly generated *COMPAS* dataset, respectively. In them, it can be observed that when balancing criminals according to race and eliminating the racial bias a minimal impact is detected, since now there is a greater number of reoffense risk with medium value, but not enough to be a negative impact. This behavior is similar for the rest of the attributes.

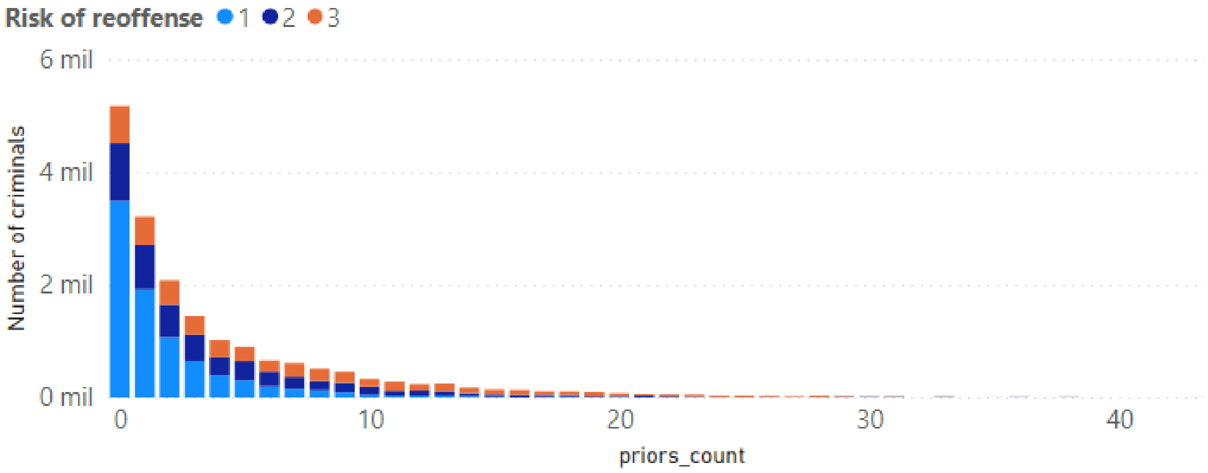


Figure 16. Risks with the unbalanced COMPAS dataset based on prior crimes

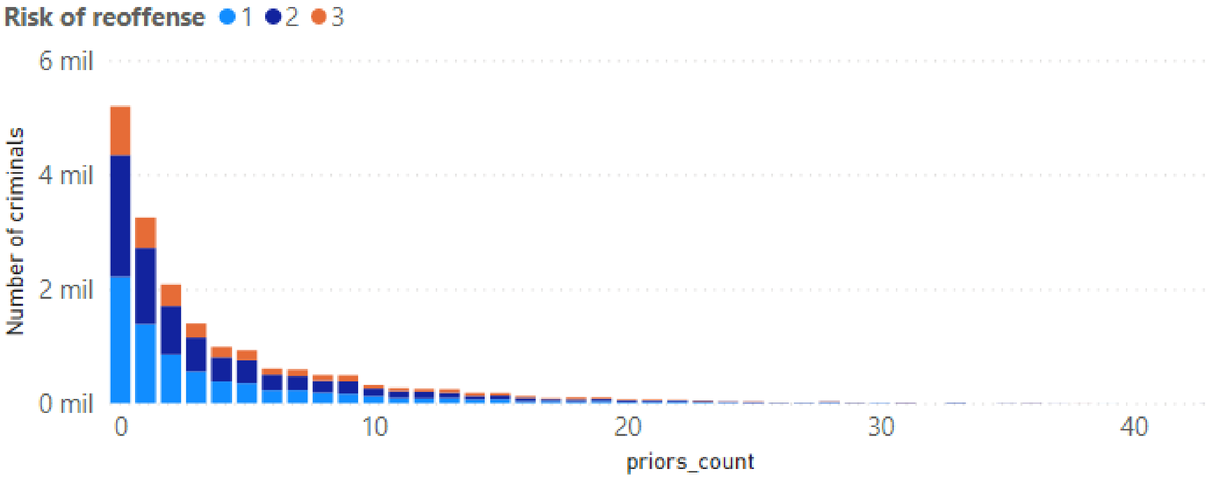


Figure 17. Risks with the balanced COMPAS-based dataset based on prior crimes

Annexes

[1] Example of a user profile CSV, the values are separated from each other by a semicolon.

```
id;unknown;Action;Adventure;Animation;Children's;Comedy;Crime;Documentary;Drama;Fantasy;Film-Noir;Horror;Musical;M
ystery;Romance;Sci-Fi;Thriller;War;Western;other
1;(+)  0.10102245062338432;(+)  0.008348879741000792;(+)  0.08933876846667874;(+)  0.17508928850404548;(+)
0.1047470717821464;(+)  0.12899906420366283;(+)  0.053467664765223764;(+)  0.10102245062338432;(+)
0.3566266301683493;(+)  0.10102245062338432;(+)  0.10102245062338432;(+)  0.06680760055560014;(+)
0.08274858479192332;(+)  0.10102245062338432;(+)  0.20011460339882337;(+)  0.2229841176513069;(+)
0.11361848035062583;(+)  0.11560482920831491;(+)  0.10102245062338432;0.0
2;(+)  0.10682732016168404;(+)  -0.05116282556558377;(+)  0.14128150451272423;(+)  0.10682732016168404;(+)
-0.08561700991662399;(+)  0.399477033348686;(+)  0.10682732016168404;(+)  0.10682732016168404;(+)
0.3994770333486859;(+)  0.10682732016168404;(+)  0.10682732016168404;(+)  0.10682732016168404;(+)
0.10682732016168404;(+)  0.10682732016168404;(+)  0.14128150451272423;(+)  0.14128150451272423;(+)
0.10682732016168404;(+)  0.14128150451272423;(+)  0.10682732016168404;0.0
3;(+)  0.033796092965125056;(+)  0.18297967666658432;(+)  0.033796092965125056;(+)  0.033796092965125056;(+)
0.033796092965125056;(+)  0.019796881124166;(+)  0.033796092965125056;(+)  0.033796092965125056;(+)
-0.052028107313706246;(+)  0.033796092965125056;(+)  0.033796092965125056;(+)  0.17545974901886774;(+)
0.033796092965125056;(+)  0.033796092965125056;(+)  0.033796092965125056;(+)  0.17545974901886774;(+)
0.01319141809298792;(+)  -0.02390350479385254;(+)  0.033796092965125056;0.0
4;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)
0.11249999999999988;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)
0.616666666666666;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)
0.11249999999999988;(+)  0.11249999999999988;(+)  0.11249999999999988;(+)  0.05833333333333304;(+)
0.05833333333333304;(+)  0.11249999999999988;(+)  0.11249999999999988;0.0
5;(+)  0.06151945500392394;(+)  0.08977496343613497;(+)  -0.005813985759810916;(+)  0.18799258652845002;(+)
0.09619740435580267;(+)  0.14306696466219057;(+)  0.03328588661451762;(+)  0.06151945500392394;(+)
0.21991628755558618;(+)  0.06151945500392394;(+)  0.06151945500392394;(+)  0.02252321701516398;(+)
0.06151945500392394;(+)  0.06151945500392394;(+)  0.06151945500392394;(+)  0.14548836978742272;(+)
0.1687284386123169;(+)  0.0423589414733593;(+)  0.06151945500392394;0.0
...
```

[2] Example of a item CSV file, representing movies

```
userID;unknown;Action;Adventure;Animation;Children's;Comedy;Crime;Documentary;Dra
ma;Fantasy;Film-Noir;Horror;Musical;Mystery;Romance;Sci-Fi;Thriller;War;Western
1;0;0;0;1;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0
2;0;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0
3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0
4;0;1;0;0;0;1;0;0;1;0;0;0;0;0;0;0;0;0;0
5;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;0;1;0;0
6;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0
7;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;1;0;0;0
8;0;0;0;0;1;1;0;0;1;0;0;0;0;0;0;0;0;0;0
9;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0
10;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;1;0
11;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;1;0;0
12;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;1;0;0
13;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0
14;0;0;0;0;0;0;0;0;1;0;0;0;0;0;1;0;0;0;0
15;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0
...
```

[3] Example of user CSV

```
userID;age;gender;occupation
1;24;M;technician
2;53;F;other
3;23;M;writer
4;24;M;technician
5;33;F;other
6;42;M;executive
7;57;M;administrator
8;36;M;administrator
9;29;M;student
10;53;M;lawyer
...
```

[4] Example of context CSV

```
userID;room_comfort;room_occupancy;weather;room_cleanliness;place_in_room;eating_allowed
1;Bad;medium;windy;Medium;Far from the screen;None
2;Good;NULL;rainy;Low;Close to the screen;Only drinks
3;Adequate;medium;sunny;Low;Center;Food and drinks
4;Adequate;medium;cloudy;High;Center;Food and drinks
5;Bad;low;sunny;High;Corner of the room;Food and drinks
6;Adequate;medium;sunny;Medium;Far from the screen;Only food
7;VeryBad;medium;sunny;Low;Corner of the room;Only food
8;Good;low;sunny;Low;Corner of the room;Only food
9;Bad;medium;sunny;Low;Close to the screen;Only drinks
10;Adequate;medium;cloudy;High;Corner of the room;Only food
...
```

[5] Example of a ratings CSV file without context

```
userID;itemID;rating
1;61;4
1;189;3
1;33;4
1;160;4
1;20;4
1;202;5
1;171;5
1;265;4
2;292;4
2;251;5
2;50;5
2;314;1
2;297;4
2;290;3
2;312;3
2;281;3
2;13;4
2;280;3
...
```

[6] Example of a customized user profile CSV

```
id;Horror;Sci-Fi;Thriller;room_comfort;room_occupancy;weather;room_cleanliness;place_i
n_room;eating_allowed;other
1;(+)  0.2;0.0;(+)  0.1;(+)  0.4;(+)  0.1;0.0;(+)  0.2;0.0;0.0;0.0
2;0.0;(+)  0.2;0.0;0.0;0.0;(+)  0.05;(+)  0.05;(+)  0.5;(+)  0.2;0.0
3;(+)  0.1;(+)  0.1;(+)  0.1;(+)  0.4;0.0;(+)  0.05;(+)  0.2;(+)  0.05;0.0;0.0
4;(+)  0.1;(+)  0.1;0.0;0.0;(+)  0.1;(+)  0.2;0.0;0.0;(+)  0.5;0.0
5;0.0;(+)  0.1;(+)  0.1;(+)  0.1;(+)  0.08;(+)  0.02;(+)  0.6;0.0;0.0;0.0
6;0.0;0.0;(+)  0.2;(+)  0.1;(+)  0.5;(+)  0.01;(+)  0.05;(+)  0.04;(+)  0.1;0.0
7;(+)  0.3;0.0;0.0;0.0;(+)  0.1;0.0;0.0;(+)  0.1;0.0;(+)  0.5
8;(+)  0.1;0.0;(+)  0.1;(+)  0.1;(+)  0.2;(+)  0.4;0.0;0.0;(+)  0.1;0.0
9;0.0;0.0;0.0;0.0;(+)  0.3;(+)  0.3;(+)  0.1;(+)  0.2;(+)  0.05;(+)  0.05
10;(+)  0.15;(+)  0.1;(+)  0.05;(+)  0.1;0.0;0.0;0.0;(+)  0.4;(+)  0.2;0.0
11;0.0;0.0;0.0;0.0;0.0;0.0;0.0;0.0;0.0;(+)  1.0
12;(+)  0.05;(+)  0.15;(+)  0.2;(+)  0.5;0.0;0.0;0.0;(+)  0.1;0.0;0.0
13;(+)  0.06;(+)  0.09;(+)  0.05;0.0;(+)  0.1;0.0;0.0;0.0;(+)  0.7;0.0
14;(+)  0.2;(+)  0.05;(+)  0.15;(+)  0.2;0.0;0.0;(+)  0.2;0.0;(+)  0.2;0.0
15;(+)  0.8;0.0;0.0;0.0;(+)  0.1;0.0;0.0;0.0;(+)  0.1;0.0
16;0.0;(+)  0.9;0.0;(+)  0.05;0.0;(+)  0.05;0.0;0.0;0.0;0.0
17;(+)  0.7;0.0;(+)  0.2;0.0;0.0;0.0;(+)  0.1;0.0;0.0;0.0
18;0.0;(+)  0.1;(+)  0.8;(+)  0.1;0.0;0.0;0.0;0.0;0.0;0.0
19;(+)  0.5;(+)  0.4;0.0;0.0;0.0;(+)  0.1;0.0;0.0;0.0;0.0
20;(+)  0.5;0.0;0.0;0.0;0.0;0.0;0.0;0.0;(+)  0.1;(+)  0.4
21;(+)  0.01;(+)  0.05;(+)  0.04;0.0;0.0;0.0;0.0;0.0;0.0;(+)  0.9
22;0.0;0.0;0.0;0.0;(+)  0.1;(+)  0.1;0.0;(+)  0.1;(+)  0.2;(+)  0.5
...
```

[7] Example of an item_scheme, representing movies

```
type=Item
number_attributes=19

name_attribute_1=Film-Noir
type_attribute_1=String
number_posible_values_attribute_1=2
posible_value_1_attribute_1=0
posible_value_percentage_1_attribute_1=98.0
posible_value_2_attribute_1=1
posible_value_percentage_2_attribute_1=2.0
generator_type_attribute_1=data.generator.attribute.GaussianAttributeGenerator
important_weight_attribute_1=true

name_attribute_2=Action
type_attribute_2=String
number_posible_values_attribute_2=2
posible_value_1_attribute_2=0
posible_value_percentage_1_attribute_2=85.0
posible_value_2_attribute_2=1
posible_value_percentage_2_attribute_2=15.0
generator_type_attribute_2=data.generator.attribute.GaussianAttributeGenerator
important_weight_attribute_2=true

name_attribute_3=Adventure
type_attribute_3=String
number_posible_values_attribute_3=2
posible_value_1_attribute_3=0
posible_value_percentage_1_attribute_3=92.0
posible_value_2_attribute_3=1
posible_value_percentage_2_attribute_3=8.0
```

```
generator_type_attribute_3=data.generator.attribute.GaussianAttributeGenerator
important_weight_attribute_3=true

name_attribute_4=Horror
type_attribute_4=String
number_posible_values_attribute_4=2
posible_value_1_attribute_4=0
posible_value_percentage_1_attribute_4=95.0
posible_value_2_attribute_4=1
posible_value_percentage_2_attribute_4=5.0
generator_type_attribute_4=data.generator.attribute.GaussianAttributeGenerator
important_weight_attribute_4=true

name_attribute_5=Romance
type_attribute_5=String
number_posible_values_attribute_5=2
posible_value_1_attribute_5=0
posible_value_percentage_1_attribute_5=85.0
posible_value_2_attribute_5=1
posible_value_percentage_2_attribute_5=15.0
generator_type_attribute_5=data.generator.attribute.GaussianAttributeGenerator
important_weight_attribute_5=true

name_attribute_6=War
type_attribute_6=String
number_posible_values_attribute_6=2
posible_value_1_attribute_6=0
posible_value_percentage_1_attribute_6=95.0
posible_value_2_attribute_6=1
posible_value_percentage_2_attribute_6=5.0
generator_type_attribute_6=data.generator.attribute.GaussianAttributeGenerator
important_weight_attribute_6=true
…
```

[8] Example of a context scheme

```
type=Context
number_attributes=3

name_attribute_1=room_comfort
type_attribute_1=String
number_posible_values_attribute_1=5
posible_value_1_attribute_1=VeryGood
posible_value_2_attribute_1=Good
posible_value_3_attribute_1=Adequate
posible_value_4_attribute_1=Bad
posible_value_5_attribute_1=VeryBad
generator_type_attribute_1=data.generator.attribute.RandomAttributeGenerator

name_attribute_2=room_occupancy
type_attribute_2=String
number_posible_values_attribute_2=3
posible_value_1_attribute_2=high
posible_value_percentage_1_attribute_2=25
posible_value_2_attribute_2=low
posible_value_percentage_2_attribute_2=25
posible_value_3_attribute_2=medium
posible_value_percentage_3_attribute_2=50
generator_type_attribute_2=data.generator.attribute.GaussianAttributeGenerator

...
```

[9] Example of item CSV representing criminals

```
itemID;sex;age;age_cat;race;is_recid
1;Male;69;Greater than 45;Other;0
2;Male;69;Greater than 45;Other;0
3;Male;31;25 - 45;Caucasian;-1
4;Male;34;25 - 45;African-American;1
5;Male;24;Less than 25;African-American;1
6;Male;24;Less than 25;African-American;1
7;Male;24;Less than 25;African-American;1
8;Male;24;Less than 25;African-American;1
9;Male;24;Less than 25;African-American;1
10;Male;23;Less than 25;African-American;0
...
```

[10] Example of context CSV representing criminal's contexts

```
contextID;juv_fel_count;juv_misd_count;juv_other_count;decile_score;priors_count;c_charge_degree;is_violent_recid;v_decile_score;v_score_text;event
1;0;0;0;-1;0;(F3);0;5;Medium;0
2;0;0;0;-1;0;(M1);0;-1;N/A;0
3;0;0;0;-1;0;(M1);0;1;Low;0
4;0;0;0;-1;0;(N/A);0;-1;N/A;0
5;0;0;0;-1;0;(N/A);0;5;Medium;0
6;0;0;0;-1;0;(N/A);0;7;Medium;0
7;0;0;0;-1;1;(F1);0;-1;N/A;0
8;0;0;0;-1;1;(F3);0;1;Low;0
9;0;0;0;-1;1;(M1);0;-1;N/A;0
10;0;0;0;-1;2;(F3);0;1;Low;0
...
```