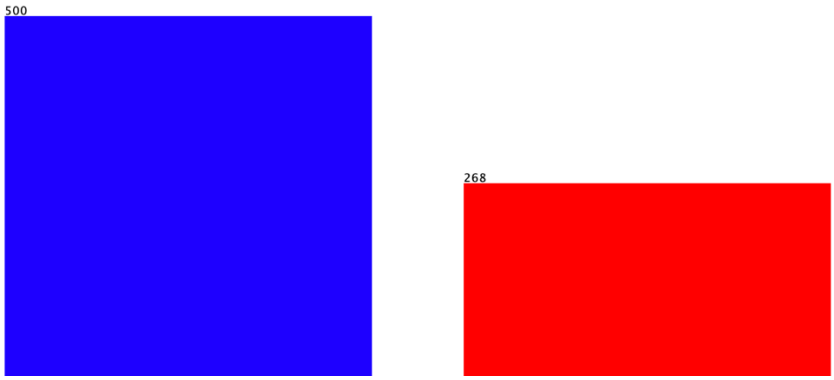
Report Diabetes dataset exercise

This dataset is composed by 768 instances corresponding to female patients.
For each patient has been considered 8 attributes:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

The patients are classified, by the 9th attribute, in positive or negative to diabetes.
The dataset is composed primary with patients that results negative to this disease, as we can see in the following diagram, in blue the negative, in red the positive.



Tasks

2 The dataset has been replaced for 20% with missing values using the filter ReplaceWithMissingValues, setting the probability to 0.2.

This are the results:

Relation: pima_diabetes

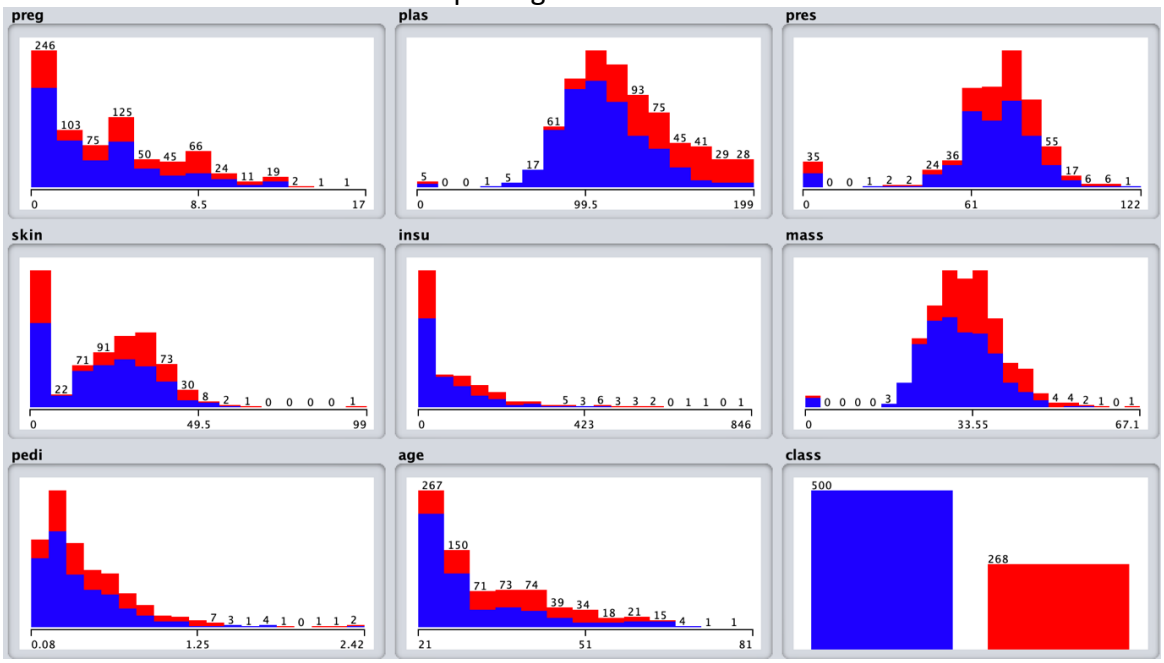
No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_negative
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested_negative
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested_positive
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_positive
25	11.0	143.0	94.0	33.0	146.0	36.6	0.254	51.0	tested_positive

Relation: pima_diabetes-weka.filters.unsupervised.attribute.ReplaceWithMissingValue-Rfirst-last-S1-P0.2

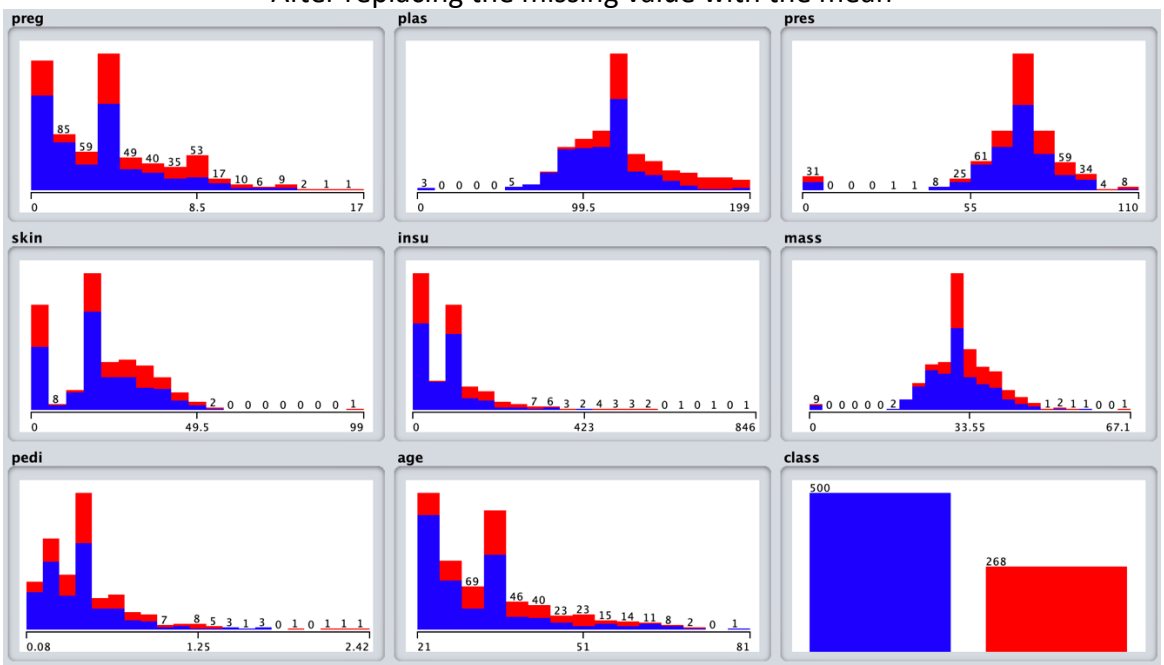
No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0		0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6		31.0	tested_negative
3	8.0		64.0		0.0	23.3		32.0	tested_positive
4	1.0		66.0	23.0	94.0	28.1	0.167		tested_negative
5				35.0	168.0	43.1	2.288	33.0	tested_positive
6	5.0	116.0	74.0		0.0	25.6		30.0	tested_negative
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0		70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0		54.0	tested_positive
11		110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12		168.0	74.0		0.0	38.0		34.0	tested_positive
13		139.0			0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0		19.0	175.0	25.8		51.0	tested_positive
16	7.0	100.0		0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0		0.551	31.0	tested_positive
18	7.0		74.0	0.0	0.0	29.6	0.254	31.0	tested_positive
19		103.0	30.0	38.0	83.0	43.3		33.0	tested_negative
20			70.0	30.0	96.0		0.529		tested_positive
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_negative
22	8.0	99.0			0.0	35.4	0.388	50.0	tested_negative
23	7.0	196.0	90.0	0.0	0.0	39.8		41.0	tested_positive
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_positive
25		143.0	94.0	33.0	146.0	36.6		51.0	tested_positive

3 The missing values has been replaced using the filter ReplaceMissingValues that substitute the missing values with the mean for the relative attribute.
As we can see in the histogram below there is an important alteration in the data.

Before replacing data with blank value

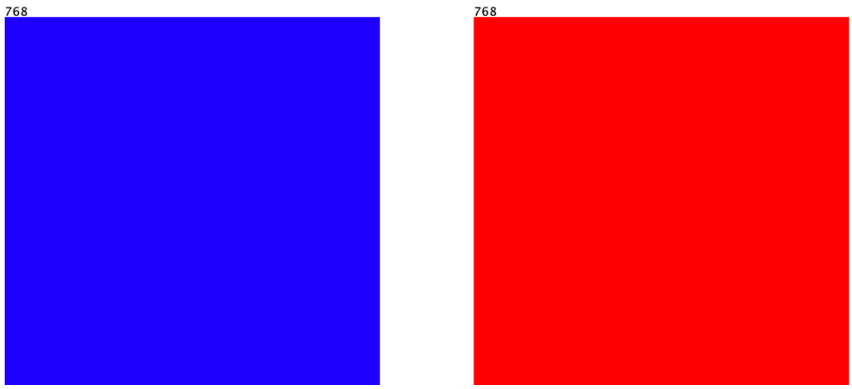


After replacing the missing value with the mean



4 The 200% resample without replacement is impossible because the instances are removed from initial space when they are sampled, so the maximum usable percentage in this case is 100%. Setting sampleSizePercentage to 200, noReplacement to false and biasToUniformClass to 1.0 we achieve the 4th task.

The instances after the sampling are 1537 as in diagram



5 To discretize data, we use the supervised filter Discretize.