**Report Diabetes dataset exercise**

This dataset is composed by 768 instances corresponding to female patients.

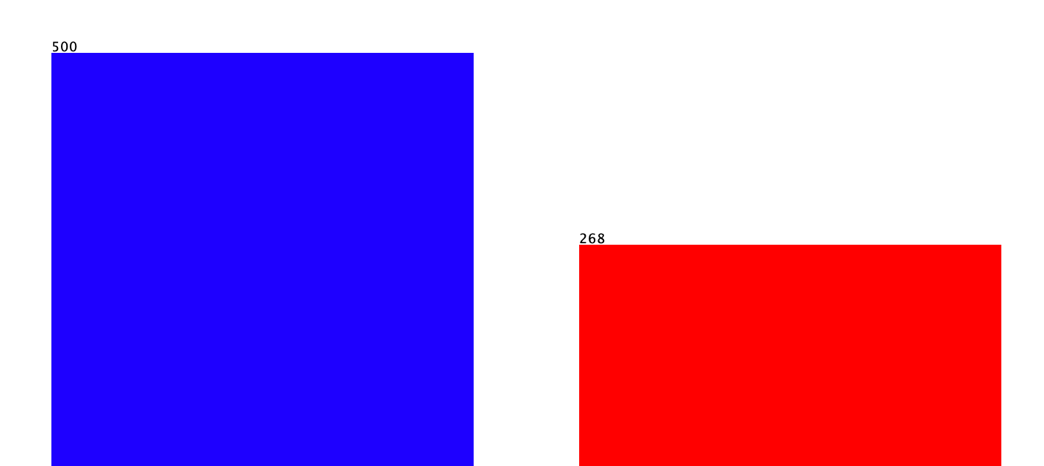
For each patient has been considered 8 attributes:

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.

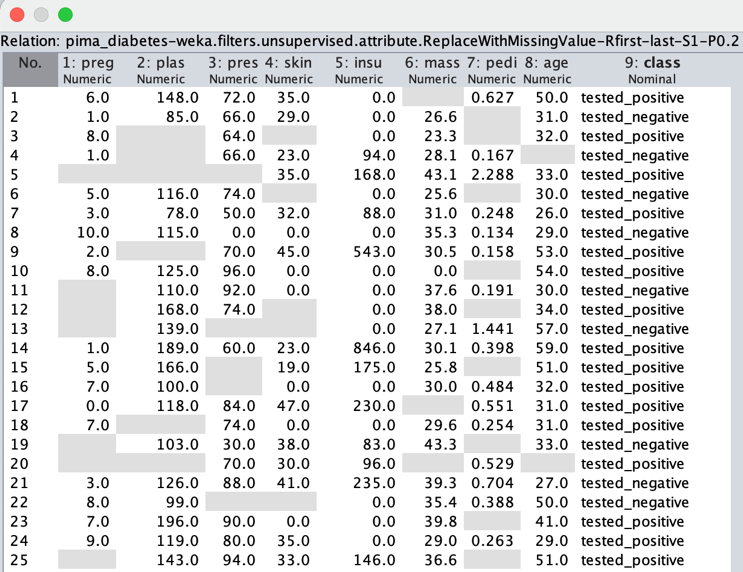


**Tasks**

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

This are the results:

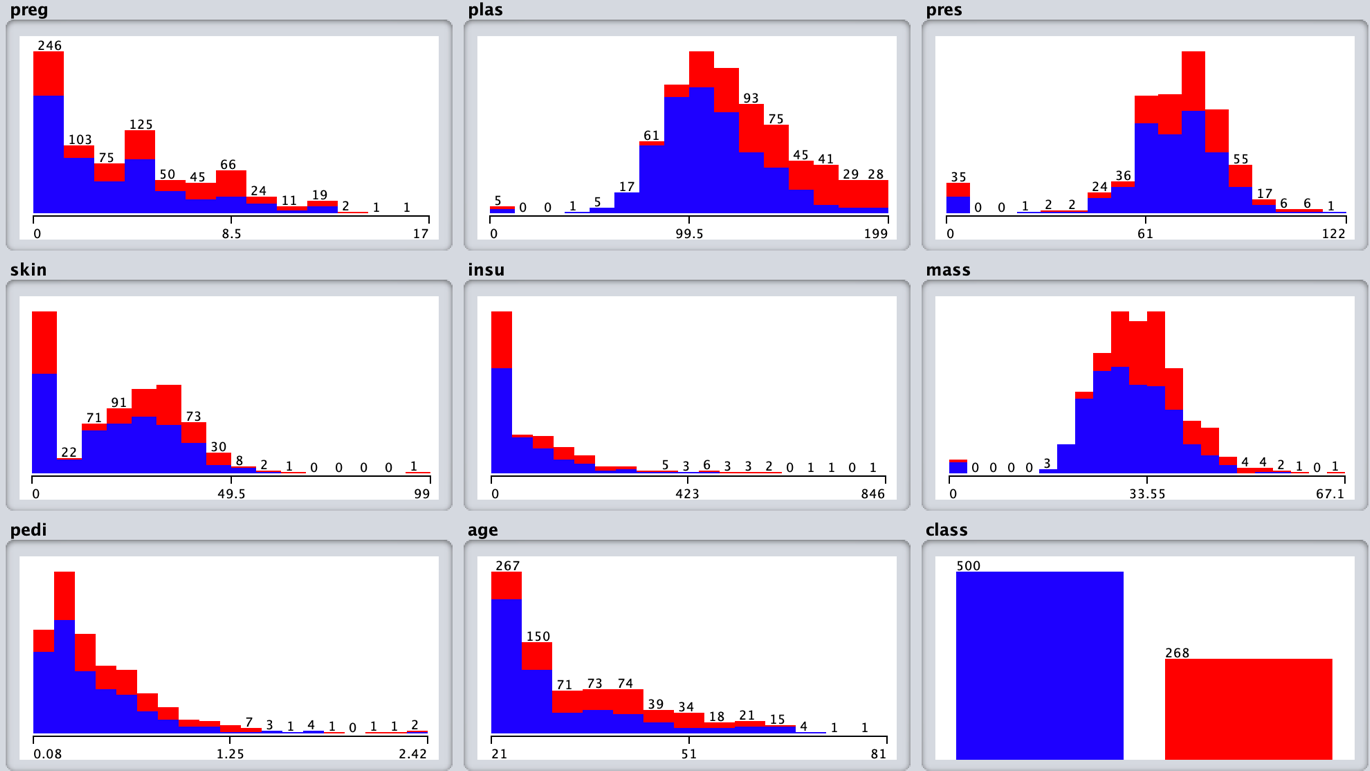
Immagine che contiene tavolo

Descrizione generata automaticamente

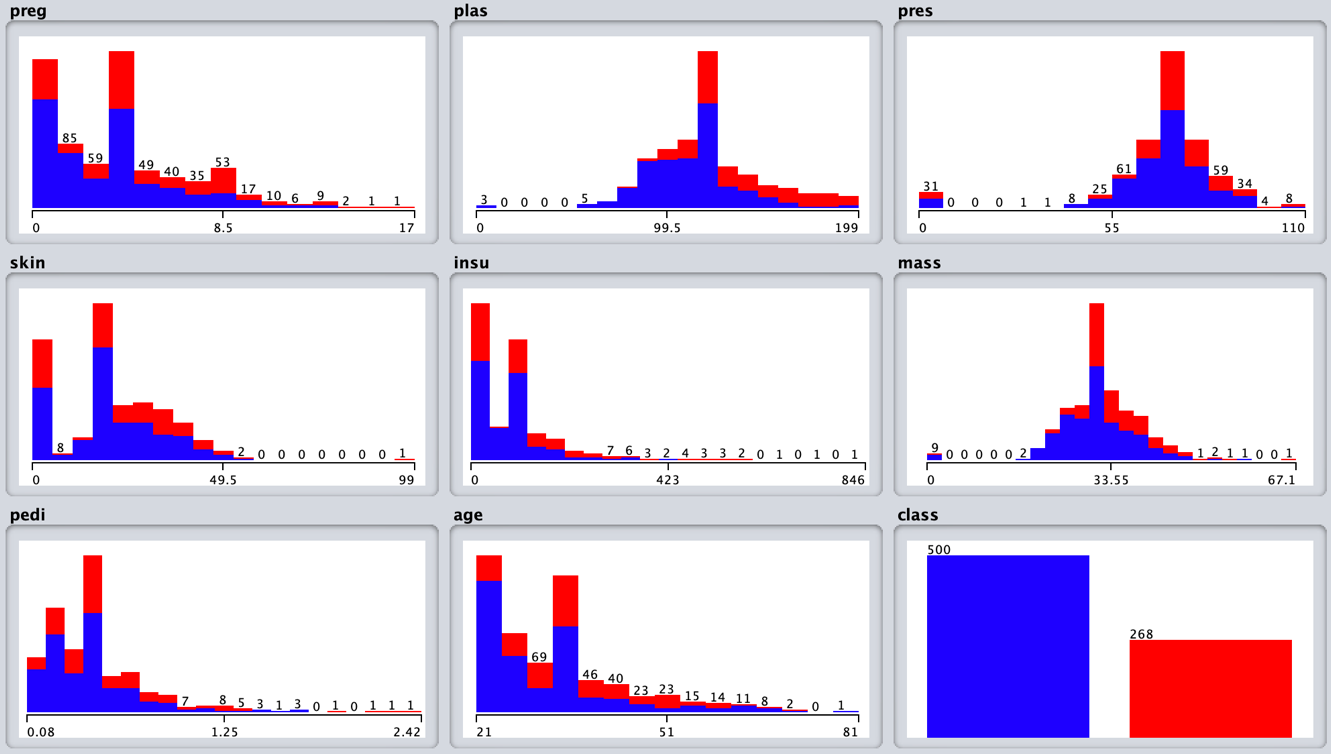
1. The missing values has been replaced using the filter ReplaceMissingValues that substitute the missing values with the mean for the relative attribute.

As we can see in the histogram below there is an important alteration in the data.

Before replacing data with blank value

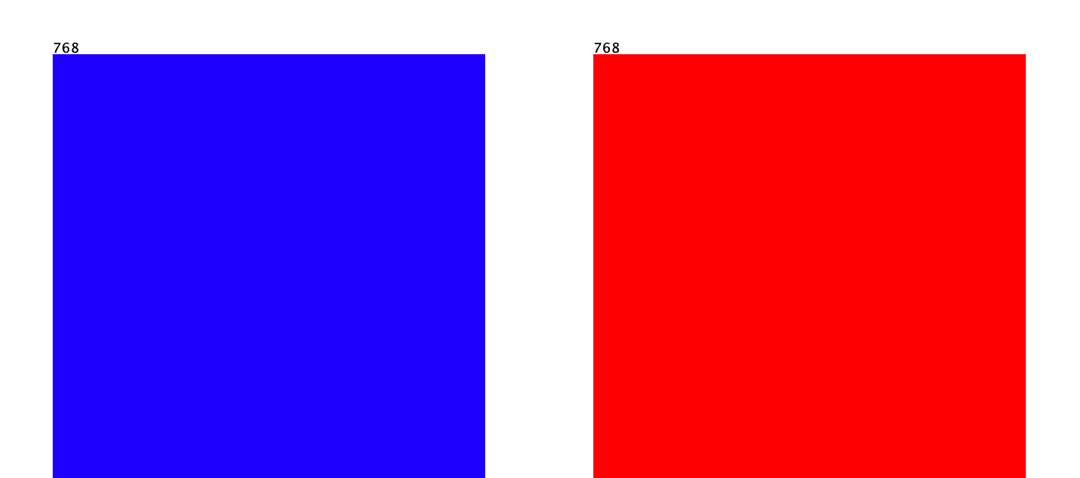


After replacing the missing value with the mean



1. 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



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

