Data Exploration and preprocessing

Section 2.2 Online Bidding advertisement

In the dataset we have both nominal and ordinal data.  
An example of nominal attribute is the Browser where is specified the browser used by the user in that observation.  
An example of ordinal data is the Inmarket\_realestate which indicates the propension of an user to bid (1) or not to bid (0) on a real estate.

In this dataset there are 25000 observations and 65 attributes.  
In the dataset there are missing values in particular the 4% of the rows has the URL attribute missing and the 38% of the rows has of the UserID missing.  
This dataset is characterized by the presence of several attributes with low information, for example the attribute Inmarket\_sportsitem has the 98.128% of the observations with value 1(yes) or the attribute Inmarket\_outdoor with the percentace above the 99%,this is solved thanks at the balancing process that we will perform before splitting the dataset.

The following columns have been removed under the assumption that they depends only on the single observation and consequently generate noise without bringing additional information.  
the columns are:'RowID', 'UserID' , 'BidID', 'IP' , 'Domain', 'URL', 'Time\_Bid', 'AdslotID'.

In the dataset the columns 'Browser' , 'Adslotvisibility' , 'Adslotformat' have been converted using a label encoder.

Section 2.3 Breast Cancer

In this dataset every attribute is a measurement done during a medical exam.  
The attributes are all numerical.  
Examples of the attributes are: concavePointsMean, symmetryMean,fractalDimensionMean, radiusStdErr, textureStdErr.  
There are no missing values and we did not preprocessed any attribute.  
Due the domain of the attributes there are no low information columns so we have used them all in order to fit the classifiers.