Cytological Reports Analysis

Stefano Valentini - m. 254825
Valentina Cecchini - m. 255596
July 2018

INTRODUCTION

This work aims to analyse a dataset containing data about cytological reports[1].

The medical exam that is object of this analysis is called “Pap test” (Papanicolaou test) and it is used to detect potentially pre-cancerous and cancerous processes in the cervix. Abnormal findings are often followed up by more sensitive diagnostic procedures, and, if warranted, interventions that aim to prevent progression to cervical cancer.

The result of this exam is composed by two “aspects”:

Cytological (visual detection of cellular lesions, main goal of this kind of test)		**Bacteriological** (visual detection of bacterial infections)
LSIL: low-degree squamous intraepithelial lesion		Trichomonas
HSIL: high-grade squamous intraepithelial lesion		Actinomyces
SCC: squamous cell carcinoma		Clamydia
ASC: atypical squamous cells	**ASC-US**: atypical squamous cells of undetermined significance	Candida
	ASC-H: atypical squamous cells, cannot exclude HSIL	…
AGC: atypical glandular cells	**AGC-NOS**: atypical glandular cells not otherwise specified	…
	AGC-neoplastic: atypical glandular cells, suspicious for AIS or cancer	…
AIS: adenocarcinoma in situ		…

Usually, the cytopathologist does not look for bacterial infections, but it reports them if they are found, moreover, it is also possible that during the analysis, traces of the HPV/HIV/Herpes viruses are found (this often leads to further laboratorial analysis).

With this work we aim to extract useful information about the characteristics that identify the subjects that are found to be affected by these pathologies; we also try to build a prediction model based on the features that are available, so to provide assistance to the cytopathologist.

[1] an example of a (redacted) report of such medical exam can be found together with this document (*docs/report_redacted.pdf*).

DESCRIPTION OF THE DATASET

The dataset has been extracted from a Microsoft Access database (single table) that belongs to a very old informative system.

The dataset is composed by 72 mixed variables and 9323 readings; the “input” variables are first 19 columns of the dataset; they provide information about the patient, its medical history and the symptoms she has manifested.

The remaining variables are used by the system to fill the report’s checkboxes, in fact, each checkbox of the medical report is represented by a boolean variable; e,g, the variable Z31_1 indicates if the first checkbox on the Z31 section is checked or not.

There is also a NOTE column that contains free text, that is used to give a description of what has been found during the analysis.

DATA CLEANING

The dataset has been subjected to an extensive cleaning and refactoring work (several columns had inconsistent values/data types, e.g. the value “01/2012” was present as “GENN2012”, “GENNAIO 2012”, etc.).

In particular, the first thing that has been done was to assign fixed types to the columns we knew contained a certain data type; for instance columns 5 to 11, 13 to 14, 19 to 27, 29 to 69 are boolean.

We also forced to “str” the column 18 to avoid parsing errors.

After a short inspection we decided (after consulting the cytopathologist) to drop the column DATI_OBIETTIVI, TIPO_OG and TIPO_TER_ORM because only a very small portion of the rows were not blank (and the column contained free text that would have been difficult to categorise).

Then, we proceeded to refactor columns that contained mixed data types, such as DATA_ULTIMA_MESTRUAZIONE.

This column contained either the date in which the subject has had her last period or the fact that the subject was in menopausa, had amenorrea or is in a post partum phase.

So we created other 3 boolean columns, named MENOPAUSA, AMENORREA and POST_PARTUM that contained a True value if the subject was in one (or more) of these state; the respective entries on the DATA_ULTIMA_MESTRUAZIONE column have been emptied.

We also decided to drop the rows that had both DATA_PRELIEVO and DATA_ULTIMA_MESTRUAZIONE to NaN and the rows that had ETA_PRELIEVO < 12 or NaN.

Other wrong values have been removed such as the rows in which DATA_ULTIMA_MESTRUAZIONE was greater than DATA_PRELIEVO, then we casted both columns to “datetime”.

We proceeded to clean the DATA_PAP_PRECEDENTE column, in fact we noticed that this column contained a lot of inconsistent values (as said at the beginning), for instance, the value “01/2012” was present as “GENN2012”, “GENNAIO 2012”, etc. and this happened for every month/year combination.

So, using regular expressions we tried to recover most of the wrongly-typed entries.

We also had entries in DATA_PAP_PRECEDENTE that contained values of the type “xx anni fa”” or “xx mesi fa” with respect to DATA_PRELIEVO, so we had to compute the date value by subtracting the right number of days from the respective DATA_PRELIEVO value.

Similarly to the DATA_ULTIMA_MESTRUAZIONE column, DATA_PAP_PRECEDENTE contained the value PRIMO_PAP in case of the subject was having his first pap test, so we added a boolean column PRIMO_PAP, removing the incorrectly placed values from DATA_PAP_PRECEDENTE.

We also added a True value to the PRIMO_PAP column for every row that had not a value in DATA_PAP_PRECEDENTE and ESITO_PAP_PRECEDENTE.

Following a consult with the cytopathologist, we decided to consider all the subjects in which DATA_PRELIEVO - DATA_ULTIMA_MESTRUAZIONE > 2 years to be in menopause.

The same situation that we had in the DATA_PAP_PRECEDENTE was present in the ESITO_PAP_PRECEDENTE column, that is: every value that could have been considered as NEGATIVO (“neg.”, “negativo”, “Negativo”, “meg.”, etc.) has been found using regular expressions and adjusted; the same procedure has been applied to the ASCUS, LSIL, etc. -like values.

Now the ESITO_PAP_PRECEDENTE contains a lot of values that belong to different categories (contains both bacteriological and cytological results), so we decided to split these categories into two new columns named CITO_PREC and BATT_PREC.

Similar regular expression-based cleaning procedures have been applied to the GRAVIDANZE and ABORTI columns.

Groups of columns were used to define that a checkbox was actually checked in the report file, so these columns have been refactored to a single column indicating the meaning of the checked value, e.g.:

…	Z31_1	Z31_2	Z31_3	…
…	FALSE	TRUE	FALSE	…
…	FALSE	FALSE	TRUE	…
…	TRUE	FALSE	FALSE	…

has become:

...	VALIDITA_CAMPIONE	...
…	SUBOTTIMALE	…
…	INADEGUATO	…
…	ADEGUATO	…

so, multiple boolean variables have been converted into a single nominal variable; this process has been applied to all the “ZXX_X” variables (and other output variables such as CONTROLLO).

The last (and most complex) column to be cleaned was NOTE.

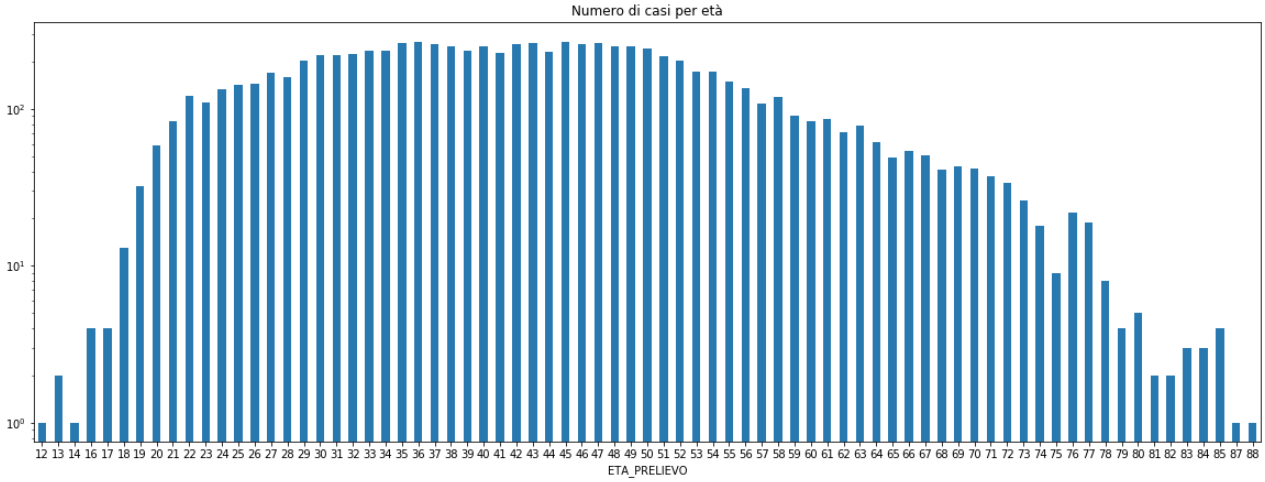
This column contained free text, that was referred to several aspects of the pap test; we had a total of 819 different values and most of them indicated the same thing.

Most of them explained the kind of bacterial infection that has been found, so we have been able add these details to the BATT column, other values referred to the CITO column, so we used this information to further fill the column.

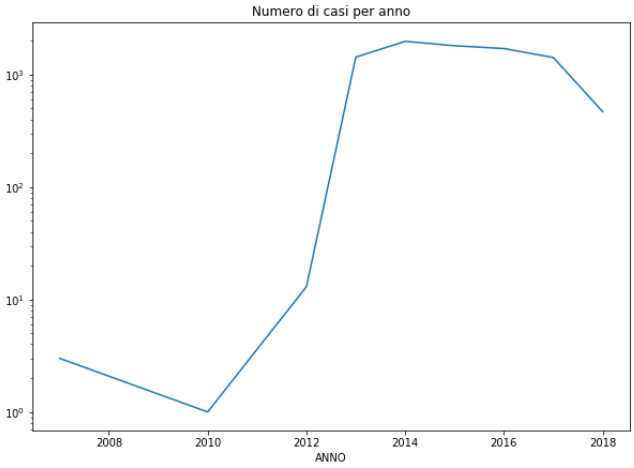
At the end of the cleaning process we had a dataset composed by 32 columns and 8804 rows.

EXPLORATORY ANALISYS[2]

We started by analysing the age of the subjects; with a mean of 42 and a standard deviation of 12.5 we can see how the population has, usually, an age that is in the range of 29-53.



Something more interesting is that the amount of the performed diagnosis has reached a peak in 2014 (the observed period goes from 2007 to 2018, even if this phenomenon is influenced by the habits and the evolution of the cytopathologist).



The number of cases is increasing in the years 2010, 2011, 2012 and 2013 as this is the time interval in which the cytopathologist started to increase its operatively.

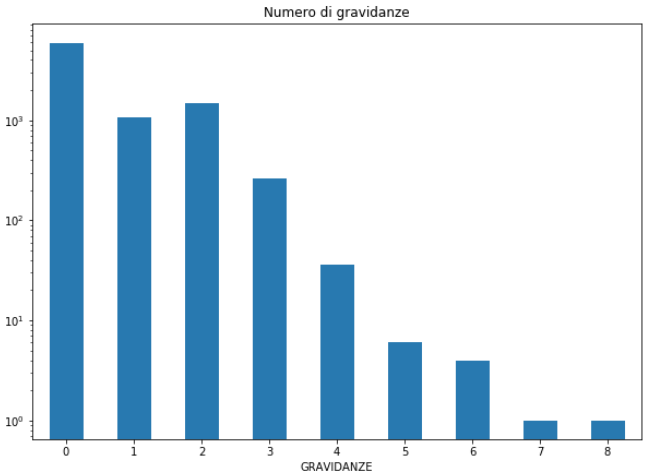
The same number of cases is decreasing in 2018: this is given by the fact that the dataset has been extracted in June 2018, so we have only the observations for the first half of the year.

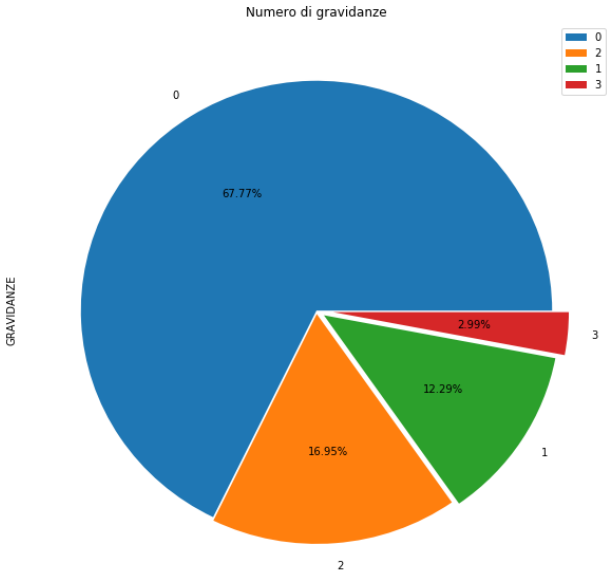
So, in conclusion, we can observe that, in the “regular” years, the number of patients is virtually the same.

[2] all the plots (even the ones that are not showed in this document) can be found in full resolution in the *plots/* folder.

Looking at the distribution of the subjects by the number of pregnancies we can see that there is an average of 0.57 pregnancies with a standard deviation of 0.92.

We can see how the majority of the subjects has had 0 pregnancies, followed by 2 and 1.





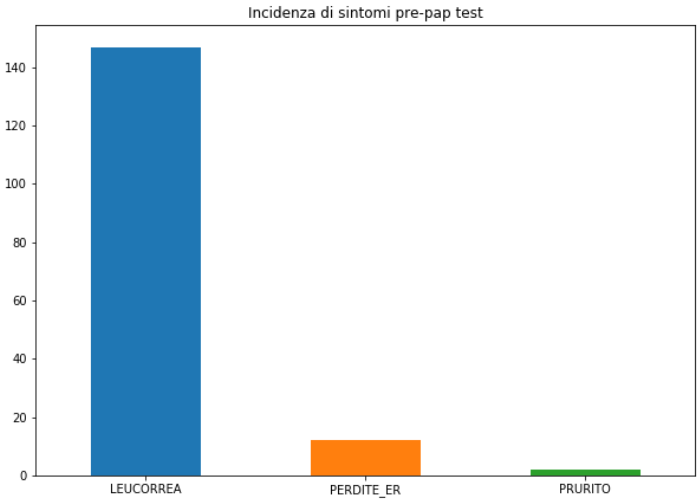
In the same way, we can take a look at the number of abortions; obviously, the majority of the individuals has had 0 abortions (91%), while the 6.78% had 1 abortions, the 1.72% 2 abortions and the 0.48% 3 abortions (a negligible number of individuals also had 4, 5, 6 and 7 abortions, while a single subject has had 8 abortions)

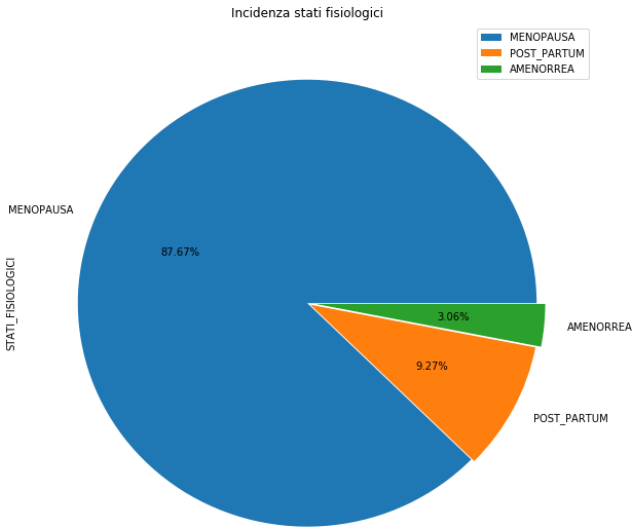
We also found that the 74.65% of the population had a its first pap test, while for the remaining 25.35% it was a subsequent check.

The organic material that is collected from the patient could not be always “readable”, in fact, we have that the 0.34% of the samples had to be re-collected.

We gathered information about the symptoms of the patient at the time of the exam by counting the “True” values in the PERDITE_ER, LEUCORREA and PRURITO columns.

We noticed that the most common symptom is LEUCORREA, followed by PERDITE_ER and PRURITO.





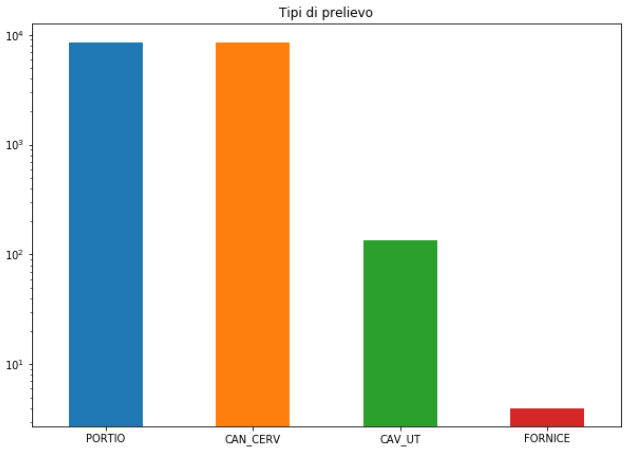
Using the same technique, we extracted information about the physiological states of the patients by analysing the MENOPAUSA, POST_PARTUM and AMENORREA columns.

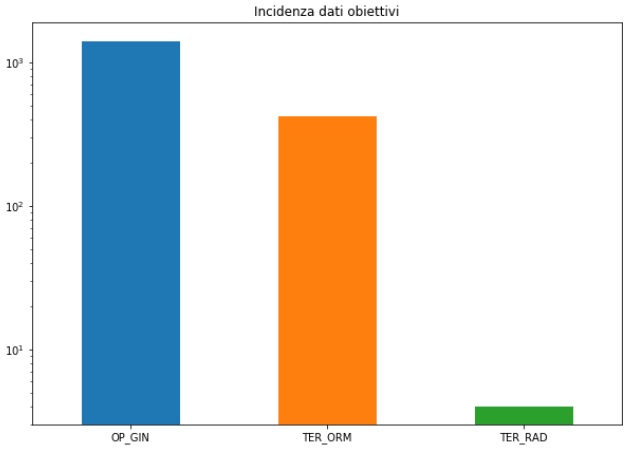
From the pie plot we can observe that the 87.67% of the population is in menopause, while the remaining 9.27% and 3.06% are in a post-partum and amenorrea state, respectively.

Another analysis that was performed by looking at several variables, was aimed to find the most common kind of biological sample collection.

We have that, usually, the collection happens in the portio - cervical canal area (even if a single sample can be collected from multiple sources at the same time)

As last “input variable” we took a look at the DATI_OBIETTIVI column.



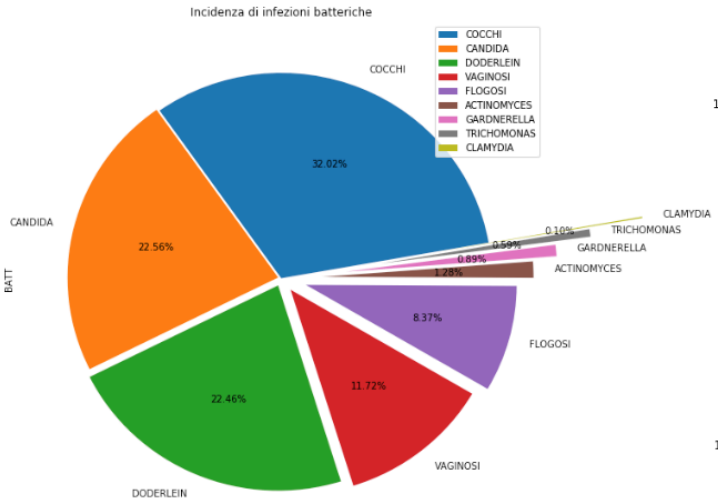


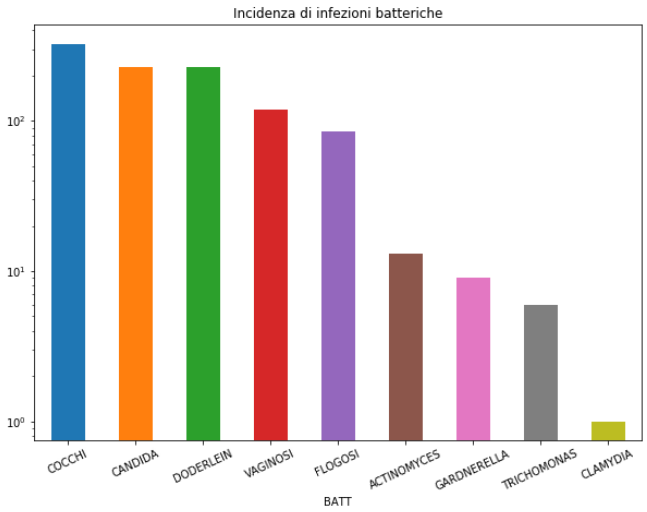
In this column we can find information about what is known regarding the patient’s medical history, such as: previous gynecological operations, hormonal therapies and radiological therapies.

From the analysis it results that the 76.77% of the population (of which individuals medical history is known - 20.83% of the entire population) has had a gynecological operation, the 23.01% has been subjected to hormonal therapy and the remaining 0.22% has been subjected to radiological therapy.

We now take a look to the “outputs” of the test.

By looking at the following plots we can observe how the population that has done a pap test is found to be affected by bacterial infections (11.52% of the total subjects):





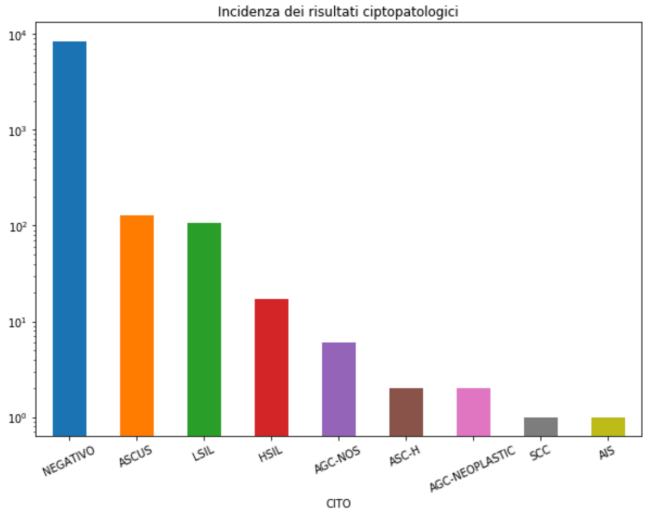
The most frequent bacterial infection is given by the coccus bacteria, followed by the candida and the doderlein.

In the same way we analysed the cytological results.

As expected, the 97.10% of the population has a negative cytological result.

By looking at scientific resources we can observe that, in fact, only the 2-5% of the population is found to be non-negative.

In the other hand, the highest positive result is represented by both ASCUS and LSIL with, respectively the 1.48% and 1.22%.



In the whole dataset we have only 1 instance of a patient affected by an SCC and 1 affected by AIS (that are the most serious and rare pathologies).

The above results have a confirmation with the data we have extracted by inspecting the variable representing the suggested time interval that has to pass until a new test is required (CONTROLLO).

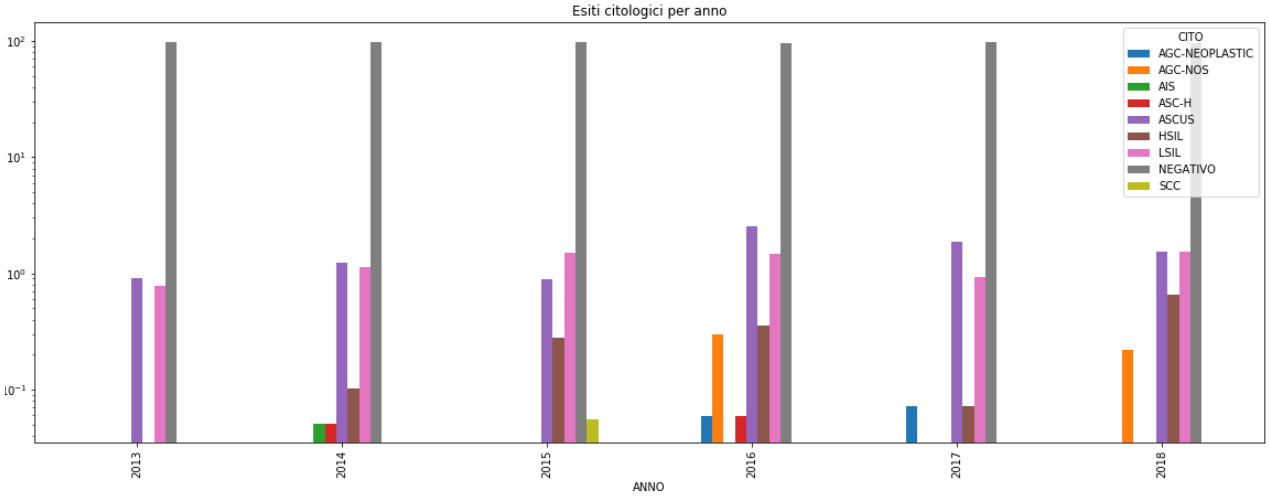
Usually, in case of a negative result we have a 12 months check, that decreases to 6 and 3 months in case of a positive result.

In fact we have that in the 98.45% of the cases a 12 month check has been suggested, in the 1.52% a 6 months one and only in the 0.02% of the cases a 3 months one.

After having analysed the single variables, we proceeded to perform some cross analysis to try to find correlations and interesting views of the topic.

All the following plots are representing percentages, to have an uniform scale.

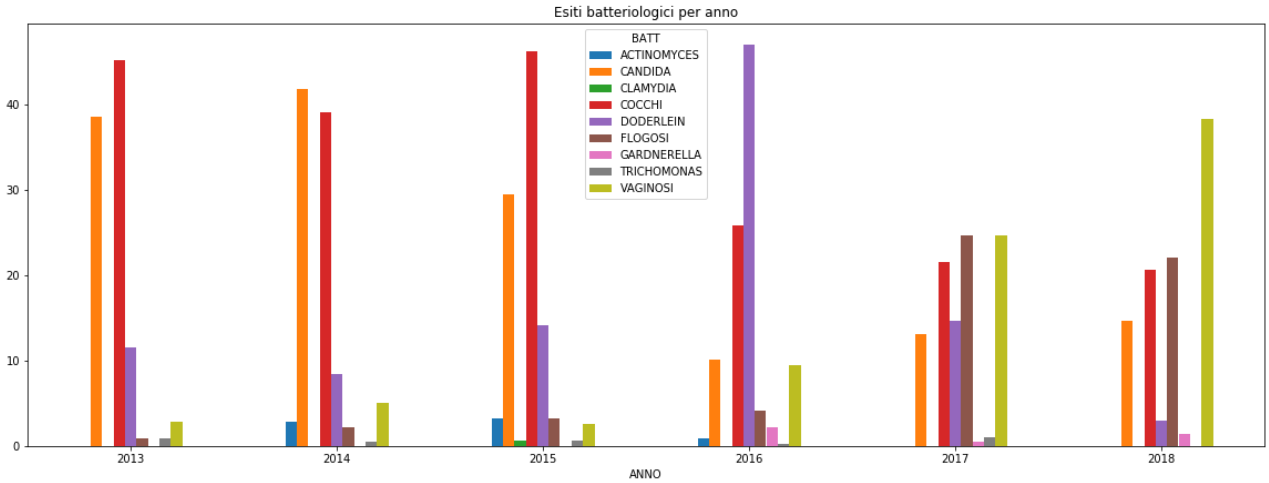
We started by crossing the DATA_PRELIEVO column with the CITO column; by grouping by the years we can have an overview of the cytological results by year.



From the plot above we can observe the following trends:

- the number of negative and LSIL results have been the same through the years (more ore less);
- we have an increasing number of ASCUS (starting from 2016);
- the number of HSIL is increasing (with a drop in 2017).

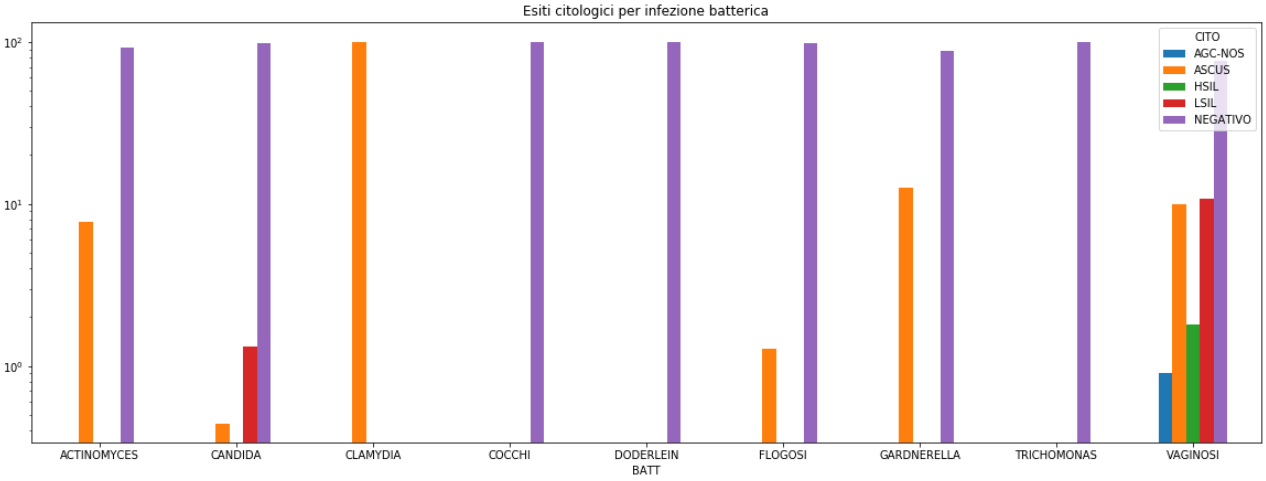
The same cross has been performed with the bacteriological results:



We observed the following:

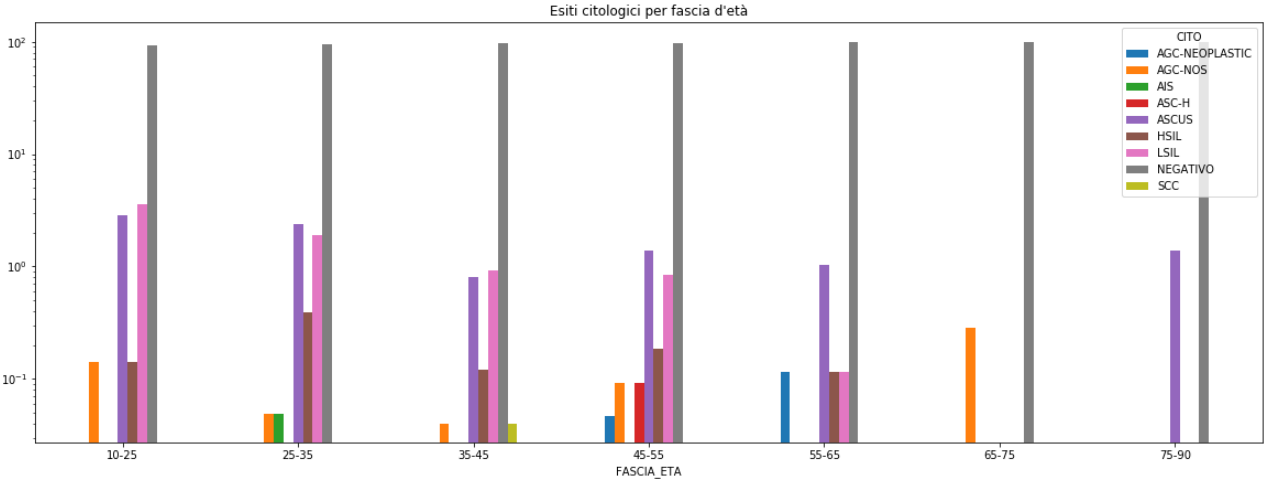
- the number of candida and cocchi infections in the patients is decreasing through the years;
- the number of doderlein infections has got a peak in 2016;
- the number of flogosi and vaginosi infections is following an increasing trend;
- gardnerella and trichomonas infections are virtually non-existing.

By crossing cytological and bacteriological results we observed the following:



- vaginosi infection does not provide a useful marker for cytological results;
- actinomyces, clamydia and gardnerella infections have an high incidence of ASCUS;
- LSIL is frequent within candida infections (even if the numbers are really not high enough to be significant).

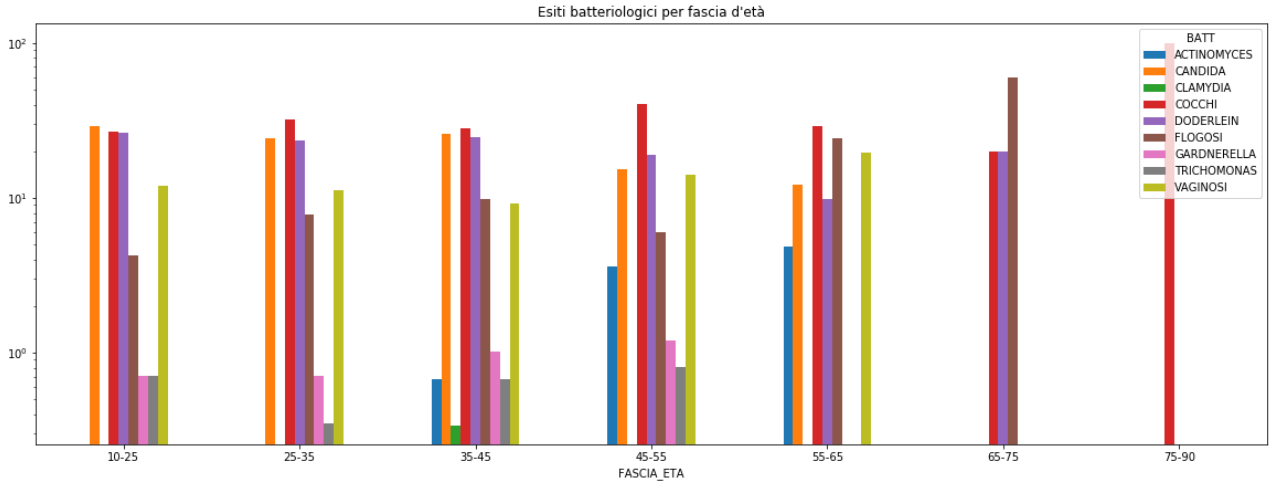
Another cross analysis has been made between the age of the subjects and the cytological results:



The ages have been binned in bins of size 10 to make the plot more readable.

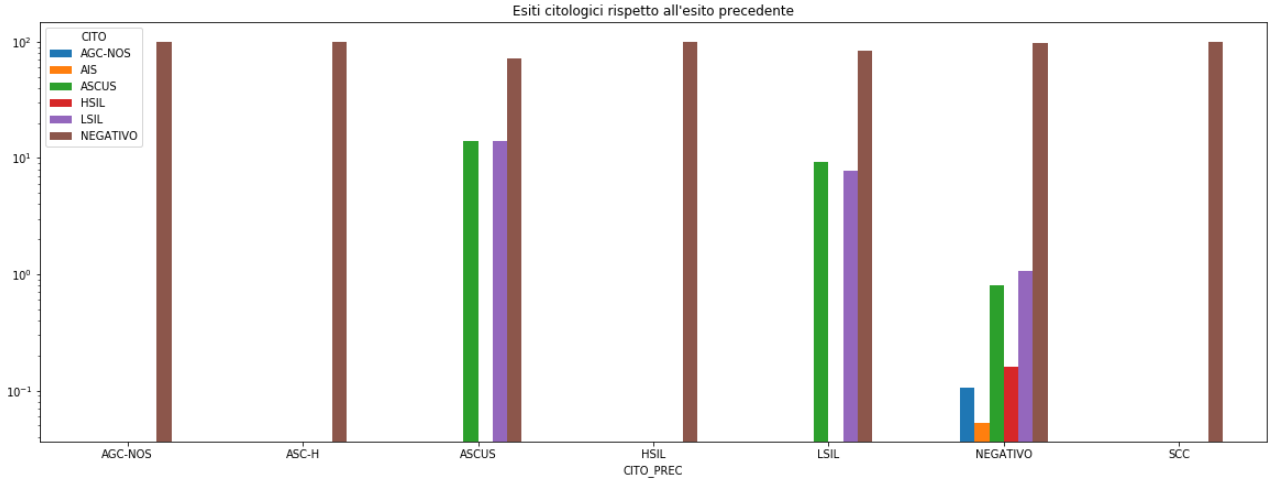
- incidence of ASCUS decreases through the years but grows again in the 75-90 range;
- LSIL are concentrated in the range 10-35; they tend to decrease with the age;
- HSIL tend to disappear after 65;
- incidence of AGC-NOS increases with the years, until 75;
- AGC-NEOPLASTIC is concentrated in the range 45-65.

In the same way, we crossed the bacteriological results:



- number of infections by actinomyces increase starting from 35, disappears after 65;
- candida infections tend to decrease as the patient ages;
- cocchi and vaginosi infections are somewhat stable, cocchi have a peak in 45-55 and 75-90;
- number of flogosi infection tend to increase;
- gardnerella increases untill 55, then disappears;

The last cross has been performed between the previous cytological result and the current one:



- patients that previously had an AGC-NOS, ASC-H, HSIL and SCC have a negative result;
- ASCUS and LSIL tend to re-propose themselves (other than becoming negatives);
- obviously, patients coming from a negative, can develop more or less any pathology.

UNSUPERVISED LEARNING

Since all the columns in the dataset can be seen are categorical (it is true that the column ETA_PRELIEVO, GRAVIDANZE and ABORTI contain integer values, but to help the clustering algorithm we decided to bin the ETA_PRELIEVO in groups, and to consider the GRAVIDANZE and ABORTI columns as True-False, or better, any value greater than 0 has been considered as True, and the values equal to 0 have been considered as false for both columns) we applied the K-modes clustering method.

The K-modes clustering is designed specifically for dataset containing categorical variables: instead of looking for the mean of variable (like the K-means clustering) it looks for the most frequent value (the mode).

These modes value will compose the centroids of the cluster, the distances from the centroids are computed by looking at the number of values in which the single individual is found to be in disagreement with the centroid.

The implementation we used its based on two criteria: *Huang* and *Cao*.

Huang's method aims to partition the objects into k groups such that the distance from objects to assigned cluster modes is minimised[3].

Cao's method selects the object with the maximum average density as the first initial cluster center; for computing other cluster centres the distance between the object and the already known cluster, and the average density of the object are considered[4].

As mentioned above the dataset has been transformed to help the clustering algorithm (already illustrated modifications are not being repeated):

- we considered only the following columns (based on the fact that they could provide useful influence in the clustering process without having to drop rows because of NaN's) ETA_PRELIEVO, GRAVIDANZE, ABORTI, CITO_PREC and CITO;
- In the CITO_PREC and CITO columns, every value that is not equal to NEGATIVO has been coerced to POSITIVO (to remove unnecessary granularity, since we have a low number of positive cases for each single pathology).

[3] https://tinyurl.com/y7jjklhc *page 7444*

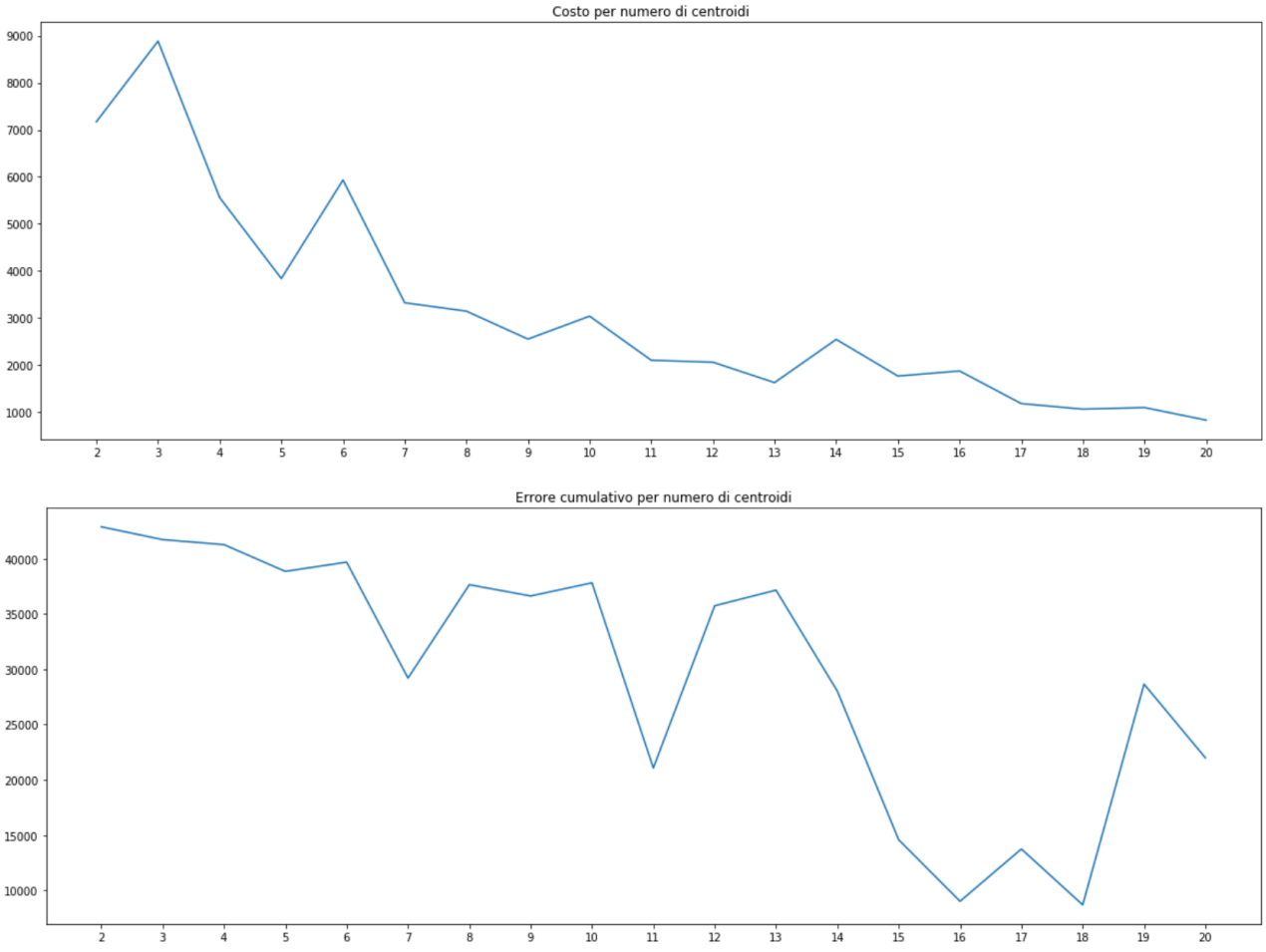
[4] https://tinyurl.com/y7jjklhc *page 7446*

We run the algorithm with both methods for a k that goes from 2 to 20.

To find the optimal value for k we used the elbow method with two different metrics:

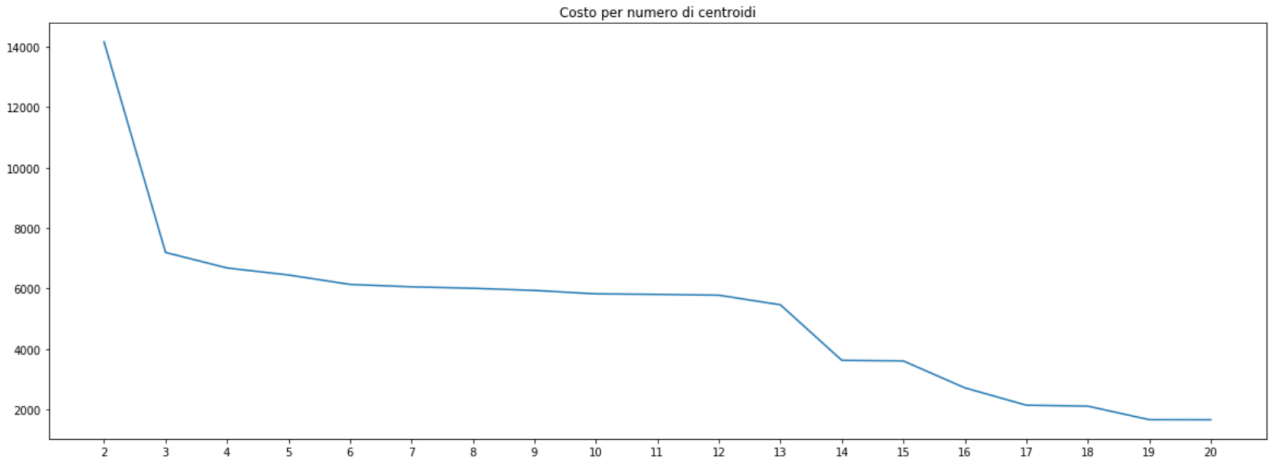
- a *cost* function that is embedded in the implementation;
- a custom *error* function (inspired by the Hamming distance) that we implemented by counting the number of differences between the individual's values and the assigned centroid; by giving more weight to errors in the CITO feature.

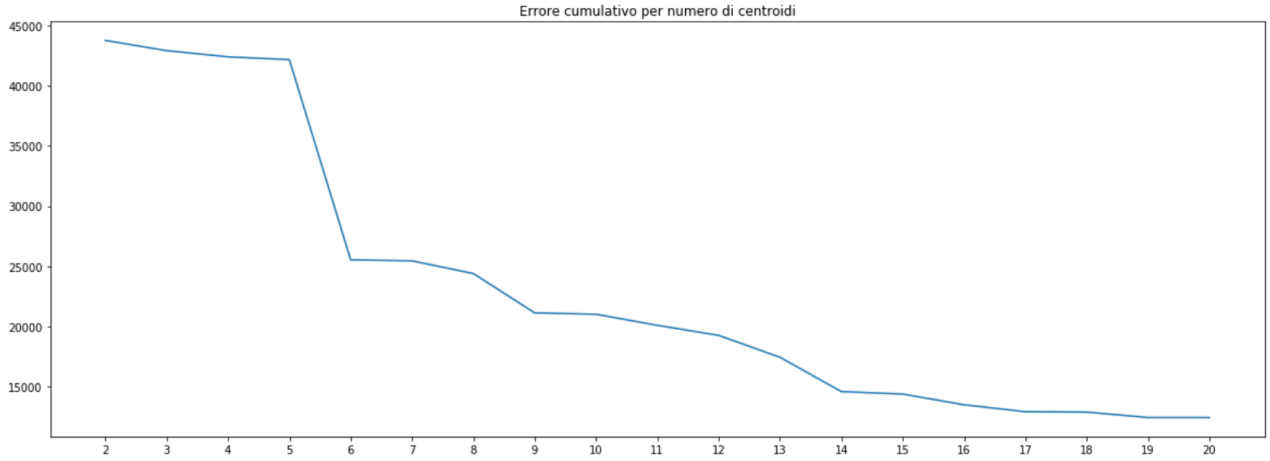
for the *Huang* method:



we can observe how the suggested number of centroids is 4, 5 or 7; clusters with size 4 and 5 have been found to be not interesting, while the cluster with 7 centroids has an interesting grouping: individuals with ETA_PRELIEVO in the range 35-55, GRAVIDANZE and ABORTI equal to False, CITO_PREC equal to POSITIVO and CITO equal POSITIVO have been grouped together.

In the same way, for the *Cao* method:





the suggested value for k is 3 or 6; the clustering with 3 centroids is not interesting, while, similarly to the previous case, the cluster with 6 centroids present an interesting association: we have the same values as the cluster described before but for the ETA_PRELIEVO value that, this time, is 10-35.

From this analysis we can elaborate that: *women with an age that goes from 10-55, with no pregnancies or abortions that have a previous positive test are more prone to be found to be positive again*.

SUPERVISED LEARNING

Using the same refactoring to the dataset applied for the clustering we tried to use several techniques to predict future pap test outcomes based on the available information.

Unfortunately, given the structure and nature of the dataset, we were not able to build an adeguate model.

More precisely, given that we want to predict the value of the CITO variable, and that all the values that can be interpreted as positive sum to 357 and the total number of observations amount to 8804, we have that the machine learning algorithm (we tried with neural networks, decision trees and logistical regression, even with cross-validation) prefer to categorise all the cases as NEGATIVO, given that the error is always insignificant.

Or better, the algorithm prefers to always wrongly categorise the positive cases, then to risk to wrongly categorise the negative cases (that are a lot more).

We even tried to balance the positive and negative cases by removing a large portion of the negative cases, but the results are still disappointing (accuracy near 50%).

In fact, even in reality, if for each new case, the cytopathologist marks it as negative, he still has an accuracy of more than 97%; that's also why trying to use a machine learning technique appears to be not feasible.

CONCLUSIONS

As pointed in the previous sections of this document, we have been able to observe several trends related to both the years and the age of the patients, interesting relations have also been found between cytological and bacteriological results, even if, maybe, the size of the dataset is not big enough to give valence to these observations.

While results from clustering have highlighted high risk categories of patients, inconclusive models have been extracted from supervised learning techniques.