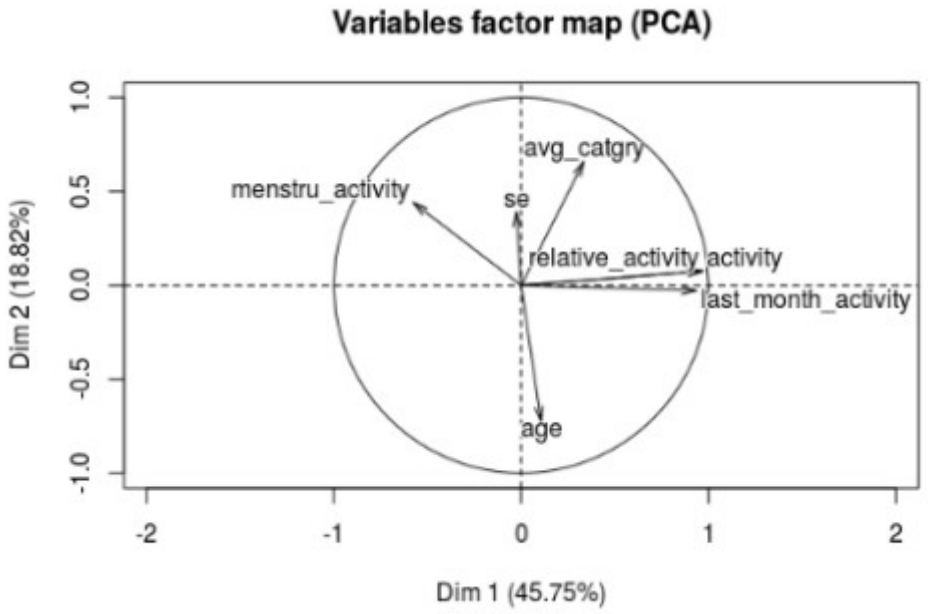
Analysis A - Principal Components Analysis



In order to cluster data in different categories to identify similar patterns we decided to use the PCA method on the variables that we have created.

The first two principal components cover almost 65% of the whole variability;

- The **first PC** represents aspects of the usage of the apps itself: it has loadings which are significant for last_month_activity, activity, relative_activity and, with opposite sign, menstru_activity. This first principal component extract information on the use of the app from the user.
- The **second principal component** on the other hand represents demographic aspects, taking in consideration variables such as age and BMI. It covers the variability coming from demography.

From the first two principal components we have used a K-means clustering algorithm (unsupervised classification) in order to identify groups of users which can have different ways of using the app. For example, if we take into consideration the first principal component, a user which uses the app frequently in the past, it is expected to have symptoms tracked more often compared to the remaining users.

Analysis B - Extracting the real probability of the symptom given the access of the user

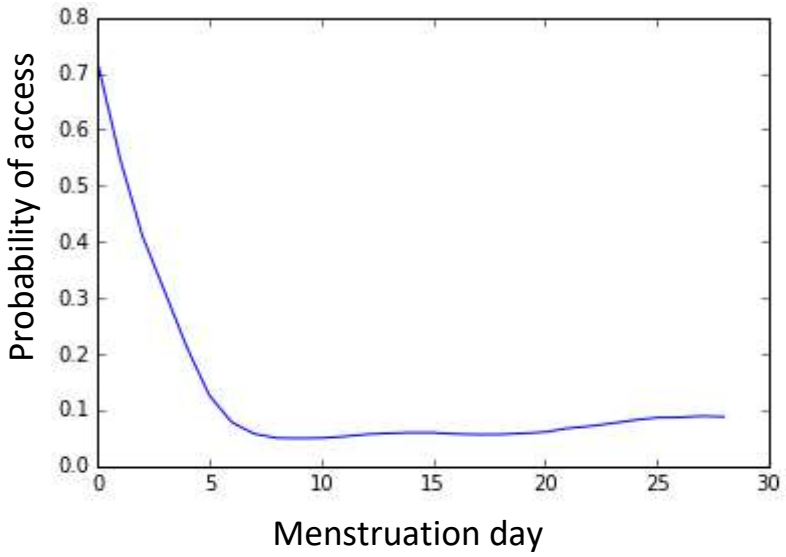
The aim of the challenge was to identify the probability

For each user, the probability of having a symptom i on the day t, it is divided in the probability of the access in the day t for the user, multiplied by the probability that the user has a specific symptom i in the day t, given that the user has made an access on that day.

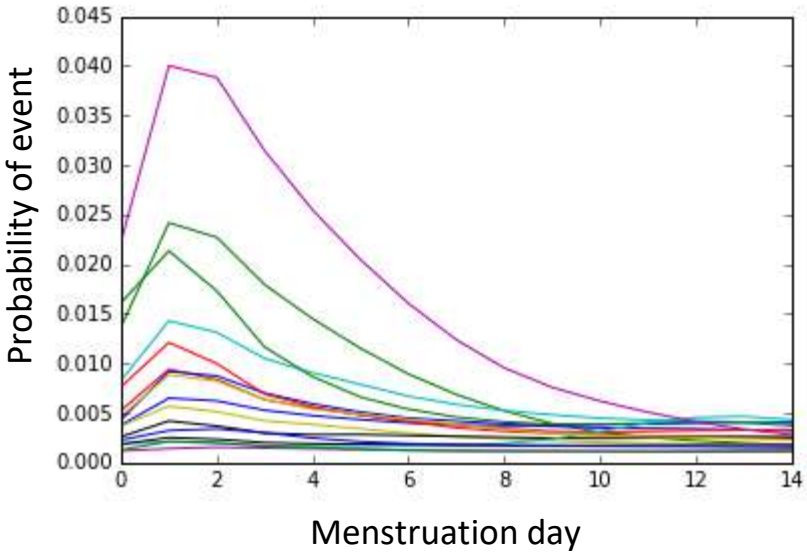
$$\Pr\left(\boldsymbol{symptom}_i^t\right) = \Pr(access^t) * \Pr\left(symptom_i^t \middle| access\right)$$

This considerations come particular useful when we talk about clustering: indeed, demographical and regional variables can influence the probability of access but not the other, and other variables affect the probability of having a symptom.

In the following figure, it is possible to see **the probability of access** for all the users (until day 29), which decays quickly after the first day of occurrence of a menstruation (most users are active in the first days). This was calculated from past available data. It is a proxy of $\Pr(access^t)$.



On the other hand, from our predicted model it is possible to retrieve the target $Pr(symptom_i^t)$. Due to the fact that our model behaves very well for the first data, but it worsen after the 15th day of menstruation, data have been reported here:



This graph is interesting, but it becomes much more interesting when we start to analyze the conditioned probability. The conditioned probability can be obtained with the following formula:

$$\Pr(\boldsymbol{symptom_i^t} \mid access) = \Pr(symptom_i^t) \;/\; \Pr(access^t)$$

In the following graph it is reported the different profiles for the different symptoms with the conditioned probability. The peaks for each symptom are shifted, showing that each symptom has different patterns across the cycle. Pink symptom is very significant for all users and it has a peak around the first day, showing a different pattern of a light-blue symptom which shows an increase in usage during the end of the period taken into consideration. With a better estimation model and the probability of access, it is possible to obtain even more significant value from the same analysis, identifying clusters.

As this is an open repository, it is not advised to mention the specific symptom but the analysis is easily reproducible with the code provided and the data.