First, we must let clear that we have decided as regression metric the mutual information, either for the most part of the lab, regression, and lastly classification, situated at the end of notebook.

After selecting 10 features from life expectancy, we perform MI metrics to both sets. We can see similarities between the ranking from each column. But, for making accurate conclusions, we decide to perform random forest regressor and represent the feature importance’s from that.

We conclude that either way, using whole data frame or with only 10 features, we should get similar or equal results due to the redundancy of the remaining columns that are not present in the smaller set. PASA DIAPO

In the third section, we decide to apply as machine learning the random forest and mutual information as regression metric. The number of components is set by a loop statement with a max range of 50, choosing as the optimal value the 19. PASA DIAPO

Here you can see the representation of all PCA components achieved. PASA DIAPO

Before applying the RF, we obtain the best hyperparameters in order to set the best scenario possible, and here are the results from grid Search while applying cross validation. PASA DIAPO

Lastly, we evaluate the model with PCA by a specific criterion:

Having all columns components

Having 19 with no countries encoded features.

With the data-frame without applying PCA we use the same criteria.

Here the tables with the results. From R-scores, we can differentiate no overfitting in any of the cases, but very tiny difference between PCA models and not. This is highly more significant if we perform other measure metrics as the lower table examples, which coincide that no PCA models are better predicting the Y values from regression than the two others. PASA DIAPO