Data preparation

For a public health organisation

Santé publique France

Yann HERENG

AI engineer training
Openclassrooms

I. DATA CLEANING PIPELINE

RAW DATA
Openfoodfacts
320772 products
162 features

STEP 01 Missing values
Removing features > 50% of missing values

STEP 02 Duplicates
Identifying primary key and removing duplicates

STEP 03 Features selection pass1
Selecting most interesting features

STEP 04 Normalisation
Normalising column headers texts dates

STEP 05 Formatting
Reassigning optimized datatype for faster computation

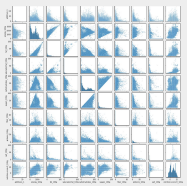
STEP 06 Outliers
Removing outliers based on quantiles and domain knowledge

STEP 07 Features selection pass2
Selecting most interesting features

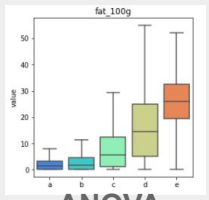
CLEAN DATA

II. ANALYSIS

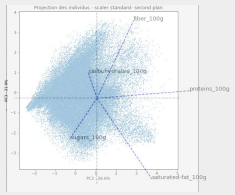
GETTING INSIGHTS FROM DATA
Using statistics and dimensional reduction



Cross-correlation

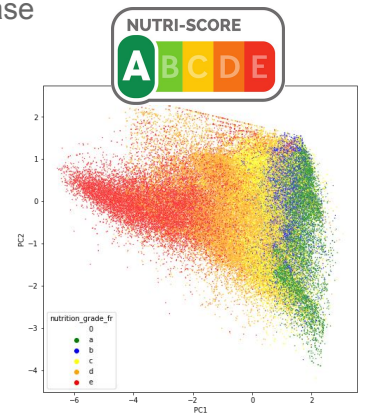


ANOVA



PCA / LDA

SUMMARIZING INFORMATION
for business use case



Automatic classification POC

TECHNICAL TOOLS USED

matplotlib · pandas · seaborn · scikit learn · plotly · jupyter · python · NumPy

github.com/yhereng/OpenClassrooms