

Project II -2020-2021 Part time

In the second project you are using the same data as for the first one. The project has two parts:

Part I

The first part aims at creating a predictive model to classify whether Clinton or Sanders will win the county.

You have to use at least 3 distinct methods and to assess how good are the predictions made by your models

Part II

The second part refers to clustering. Forget about the elections and the votes. You do not have to use them any further. Here we want to use the "economic related" variables to cluster the counties. You can use (select among) them.

The "economic related" are all the variables except from

PST045214, PST040210, PST120214, POP010210

AGE135214, AGE295214, AGE775214, SEX255214

RHI125214, RHI225214, RHI325214, RHI425214

RHI525214, RHI625214, RHI725214, RHI825214

POP715213, POP645213, POP815213, EDU635213

EDU685213, VET605213

This set is called "demographic" and you have to use them to describe the clusters you have found.

Deliverables: Provide a report in pdf format with your findings. Be as detailed as possible so as your report to be self-explained. Explain the methodologies used, how good you think it is and any limitations that may apply. You have also to upload in a separate file the R code you have used for the project.

