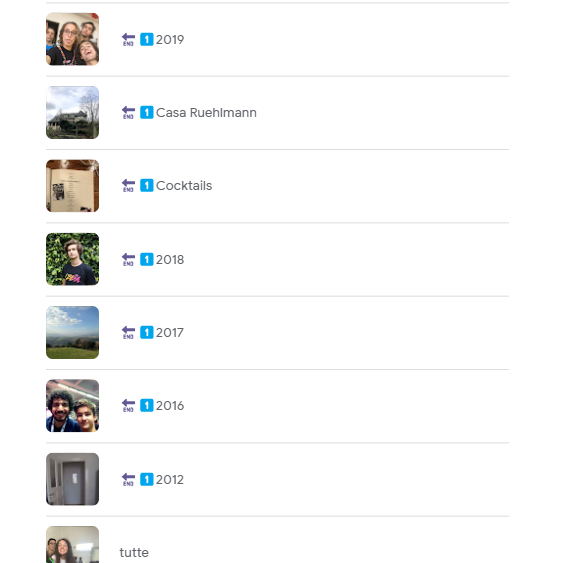
riprendere photo da venezia fatta.



mettere parigi

DOMANDE PER MIKE

* MODEL PIPELINE
  + che cazzo è? come mai la faccio prima del fit?
  + a che serve sceglie come sampler il randoOverSampler se sotto ho diversi over-sampling e dimensionality reduction method?
* GRAFICI #1
  + perché f1-score scende da 1 a 0.85? overfitting??
  + come mai il training set size è solo tra 0 e 500? non dovrebbe essere di più?
* GRAFICI #2

may not be generalizable to other datasets or future elections.  
  
Here are some reasons why it may be difficult to determine if a county will vote for a democrat based on the features of your dataset:

The features in the dataset may not be comprehensive enough to fully capture all the factors that influence voting behavior in a county.

The features in the dataset may not be relevant to all counties, and therefore may not be able to accurately predict voting behavior in all locations.

The relationship between the features in the dataset and voting behavior may be complex and non-linear, making it difficult to accurately predict voting behavior using a simple model.

The dataset may not be representative of the population of all counties, which could lead to inaccuracies in the predictions.

The dataset may not be able to capture the current political climate and its effect on voting behavior.

The dataset may not be able to capture the changes in voting behavior due to demographic shift or other factors.

It's important to keep in mind that voting behavior can be influenced by a wide range of factors, including economic conditions, social issues, and political campaigns. Additionally, it's possible that there may be other features that could be useful to include in the dataset in order to improve the accuracy of predictions.

----------------------

Yes, that's a good point. The dataset may not take into consideration the different ethnicities and personal backgrounds of the voters, which can have a significant impact on voting behavior. Some examples of how ethnicity and personal background may influence voting behavior include:

Ethnicity: Studies have shown that certain ethnic groups, such as African Americans and Hispanic Americans, are more likely to vote for democratic candidates. This could be due to a variety of factors, including historical experiences with discrimination and economic disparities.

Personal background: A voter's personal background, such as their education level, income, or occupation, can also influence their political beliefs and voting behavior. For example, a voter with a higher education level may be more likely to vote for democratic candidates because they align with the party's stance on issues such as healthcare and education.

Religion: A voter's religious beliefs can also influence their political beliefs and voting behavior. For example, a voter who is religious may be more likely to vote for a republican candidate because they align with the party's stance on issues such as abortion and same-sex marriage.

It's worth noting that these are just a few examples and that there can be many other factors that influence voting behavior, and that the relationship between these factors and voting behavior can be complex and vary depending on the location.

---------------------

may not take in consideration pethicnicty and personal background

I want to focus on the <code>f1\_score</code> that ranges between 0.75 and 0.79. This algorithm is able to predict with an f1\_score around 0.77 if a county voted in the election for Clinton (a member of the Democratic party) or not. I want to acknowledge the fact that the result is based on some objective data (like the mean age, population density, and percent of people in a nursing homes) while there are some fundamental and more personal features that are not present in the dataset.

As an example, the dataset is not taking into account the different ethnicities, religions, and personal backgrounds of the voters, which can have a significant impact on voting behavior.

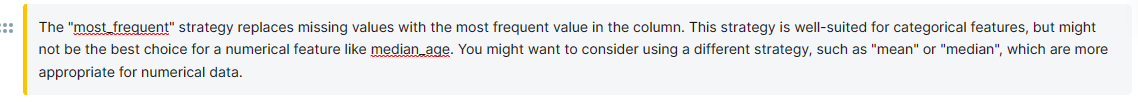
- **\*\*Ethnicity\*\*** can particularly influence the voting behavior of some ethnic groups. African Americans and Hispanic Americans are more likely to vote for Democratic candidates for different reasons (including experiences with discrimination and economic disparities).

- About **\*\*religion\*\*** we can imagine that a voter's religious beliefs will influence their political behavior. As an example, somebody who is religious might vote for a republican candidate because is aligns with the party's stance on issues such as abortion, legalization, and same-sex marriage.

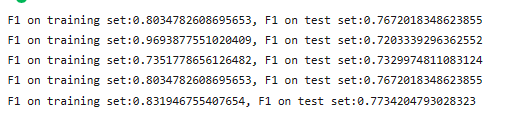
- **\*\*Personal background\*\***, on the other hand, including educational level and daily experiences. This is particularly personal because it contains the environment around a person and how the environment is actually perceived by the individual.

Those are just some examples of why I believe that predicting if a county voted for democrats or not based on those features will almost never be perfect. We have to acknowledge that people, even if they can be clustered into different groups based on income or gender there are some personal factors that will influence if they will vote for democrats or not.

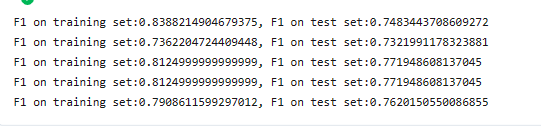
The "most\_frequent" strategy replaces missing values with the most frequent value in the column. This strategy is well-suited for categorical features, but might not be the best choice for a numerical feature like median\_age. You might want to consider using a different strategy, such as "mean" or "median", which are more appropriate for numerical data.



**FULL DATASET**



**NOT FULL DATASET (BY ME)**

****

**pipeline\_median\_age = Pipeline([**

**("1", SimpleImputer(strategy="mean")),**

**("2", StandardScaler())**

**])**

**pipeline\_mean\_savings = Pipeline([**

**("1", KNNImputer(n\_neighbors=10)),**

**("2", StandardScaler())**

**])**

**pipeline\_percent = Pipeline([**

**("1", KNNImputer(n\_neighbors=5)),**

**("2", MinMaxScaler())**

**])**

**pipeline\_state = Pipeline([**

**("1", SimpleImputer(strategy='most\_frequent')),**

**("2", OneHotEncoder(categories='auto',drop='first',handle\_unknown='ignore'))**

**])**

**pipeline\_vote\_for\_dems\_before = Pipeline([**

**("1", SimpleImputer(strategy='most\_frequent'))**

**])**

**pipeline\_general = Pipeline([**

**("1", KNNImputer(n\_neighbors=5)),**

**("2", StandardScaler())**

**])**