ML4Science

Week 1 meeting

R. Brioschi, G. D'Angeli, F. Di Gennaro

Tutor: Dr. Michele Bianco

EPFL, Laboratory of Astrophysics (LASTRO)

21/11/2022

EPFL

Outline

1. IRIS analysis on our dataframe
   - Manipulation of the dataset
   - Pairplot
2. K-Means
3. Plots with PCA

Manipulation of the dataset

- Manipulation of the dataset:
  1. Feature engineering (Velocity of Halo, Position of Halo).
  2. Discretization of M_{HI} through quantile division (6 classes were created).

(a) $z = 2$

(b) $z = 0.95$

Figure 1: Categorization of data with different z and same parameters

Pairplot

- We created the pairplot of our features:
  1. We firstly removed the features 'MetalsGas' and 'MetalsStars' due to the fact that they maintain the same value.

Figure 2: plot with $z = 0.95$ of the variables we decided to remove

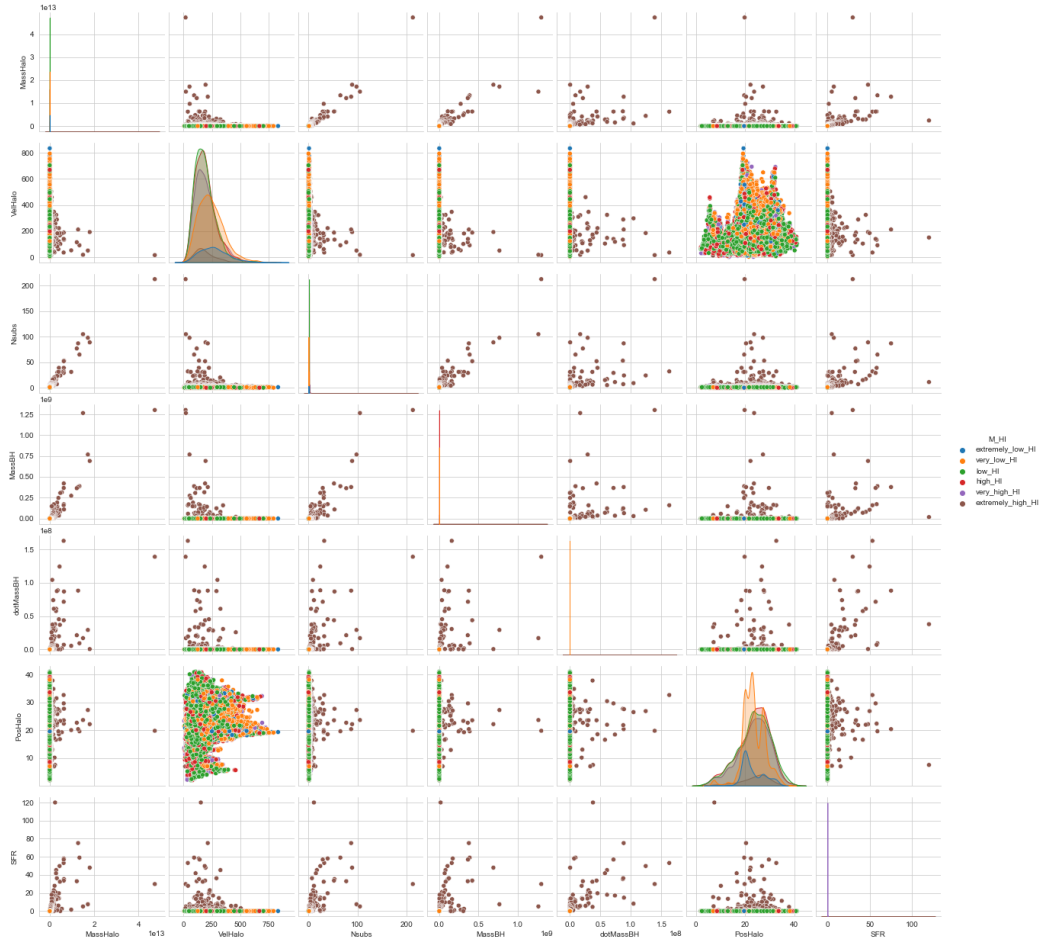


Figure 3: Pairplot with $z = 0.95$

K-Means

Supposing to not know the actual value of M_{HI}, we want to use an unsupervised learning algorithm to label each data point. We expect that this classification returns a categorization which is different from the one made by us. The steps followed were:

1. Removal of MetalsGas, MetalsStar, PosHalo and VelHalo, since they do not seem to carry useful information
2. MassHalo rescaled since its values are far higher than the others. If we do not do so, the method will overoptimize w.r.t this only variable.
3. Rescaling of the other mass values so that we can consider only their magnitude.
4. using a 1-hot encoding to deal with Nsubs is not efficient (the dimension of the feature space grows and we suffer from the Curse of Dimensionality)
5. Visualize of elbow curve to choose the optimal K (hyperparameter).

Non standardize and non rescaled data

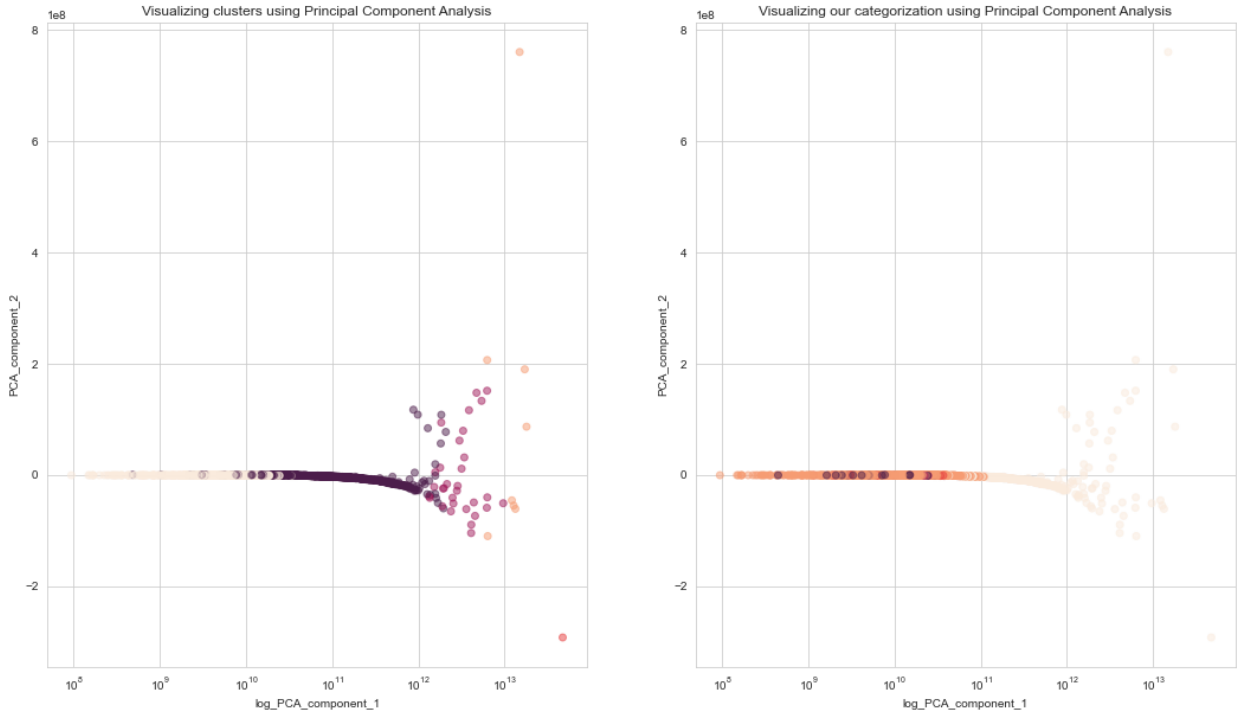


Figure 4: plot with $z = 0.95$

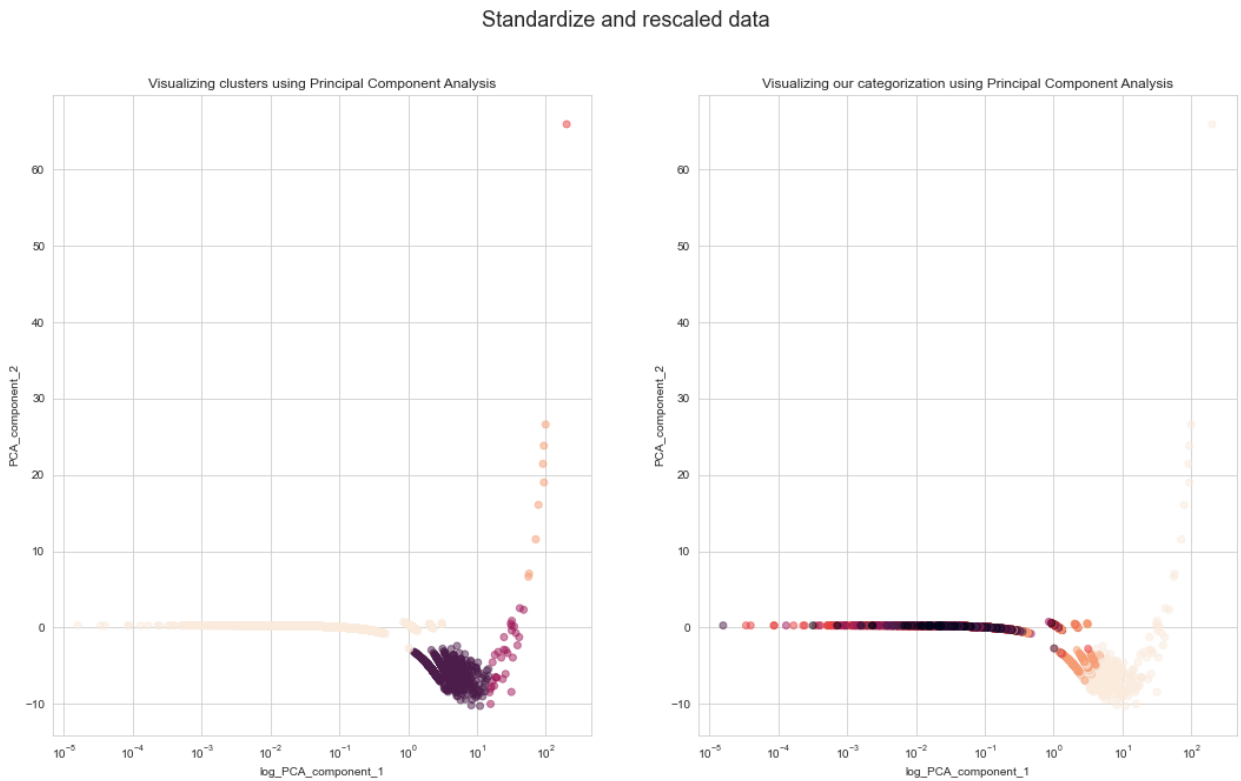
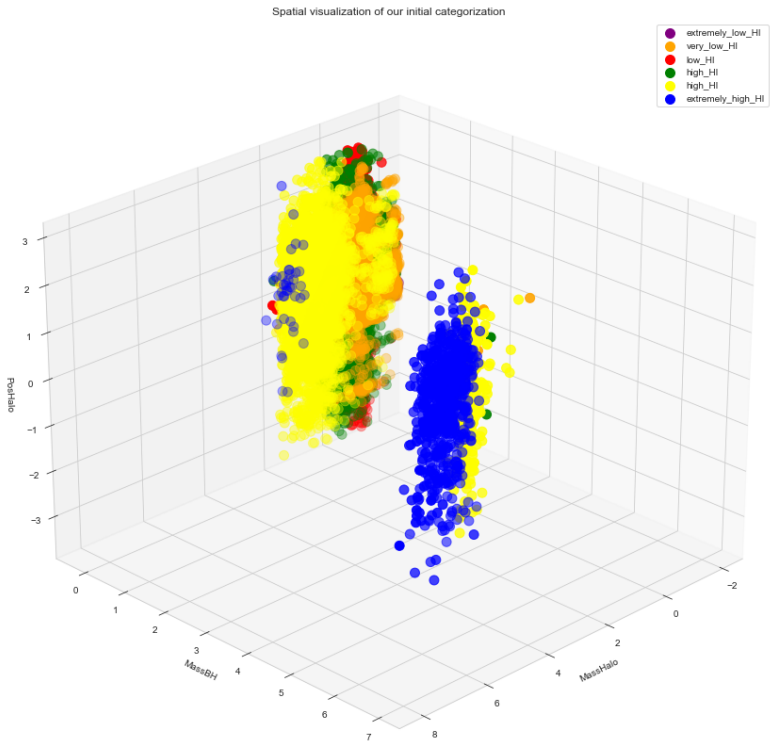
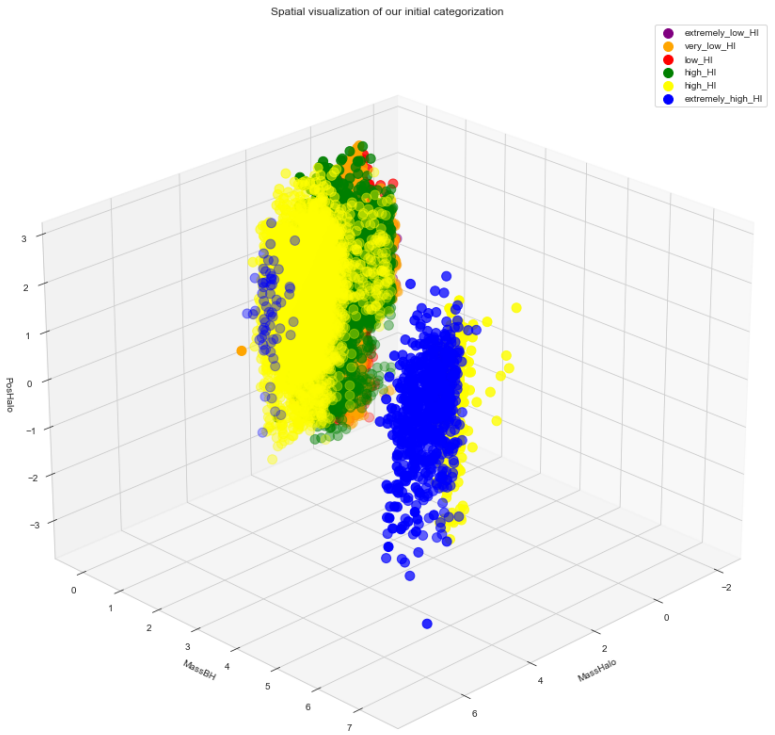


Figure 5: plot with $z = 0.95$



(a) $z = 0.95$



(b) $z = 2$

Figure 6: Spatial visualization with different z and same parameters

Conclusions

- We can conclude that the features we decided to use revealed to be good explanatory variables. When working with the final architecture (Fully Connected Neural Network), we will decide whether to keep or to drop the features we classified as meaningless.
- The last plots clearly show how the partition we made before does not only depend on the mass of the halo but also on some other relations. The other features do not introduce any other particular division in our data.