Photometric Identification of Infrared Sources

Machine Learning Project* Phase 1: **Data Preparation**

April 2, 2021

1 Introduction

Some stellar sources contain excess infrared radiation, which indicates the presence of dust around them. Dust may exist around young stellar sources (YSOs), that is the dust remaining from the star formation phase. It may also form at the final stages of a star's evolution around it when the star cools sharply and begins to produce dust due to the expansion of its outer layers. The latter stellar phase is called AGB and post-AGB. Each of these two stages has its own classifications, but what matters to us is separating young stellar objects (YSOs) from stars in their final stages of evolution (AGBs and post-AGBs). The former is a measure of star formation rate (SFR) in the galaxy, and the latter is a measure of stellar feedback and the entry of dust and metal-riched gas into the interstellar medium (ISM). Infrared sources are usually identified individually using spectroscopic data. But our goal is to see how accurately such sources can be identified using photometric data and through machine learning. For this study, we need data at different wavelengths —especially infrared. So we can use near-infrared surveys conducted by Spitzer, Gaia, WISE, etc. Finally, we will classify the sources into different categories.

*Group members: Mobin Saidy, Reyhane Javadi, Amirhosein Masoudnezhad, Negin Khosravaninezhad

2 Data

What we need is spectral data from different sources at different wavelengths. The sources' radiation at different wavelengths are our data *features*. And the kinds of sources are our data *labels*, that are given in Table 1. As you see, AGBs are themselves classified into O-rich or C-rich types based on chemistry of their photosphere and outer envelope. S stars are generally regarded as intermediate between O-rich and C-rich stars in their properties.

Table 1: Data Labels

Category	**Label**	**Encoding**
YSO	YSO	[0, 0, 0, 1, 0]
AGB	O-rich AGB	[0, 1, 0, 0, 0]
	C-rich AGB	[1, 0, 0, 0, 0]
	S-rich AG	[0, 0, 1, 0, 0]
post-AGB	post-AGB	[0, 0, 0, 0, 1]

We searched in IR surveys from different telescopes and gathered labeled spectral data. Some surveys had their features as *fluxes* for some sources and as *magnitudes* for some other. So we had to transform some fluxes to magnitudes and some magnitudes to fluxes in order to have uniform forms of features. Then we began pre-processing our data in 5 steps as follows.

2.1 Encoding

Our first step toward per-processing was to create dummy variables in order to put our categorical labels in numeric variables. We used One-Hot Encoding [1] method to create 5 variables (y_1, ... , y_5) each with values either 0 or 1 to represent our categories. You can see each category's encoding in Table 1.

2.2 Imputation

Next we had to replace the features that had NaN values with appropriate values, or remove features containing too many NaN values. For this purpose, we used Multivariate Imputer [2] that estimates NaN features from the values of other features. At last, we came to 22 features with numeric values for all IR sources. They are shown in Table 2.

[1] From scikit-learn library

[2] From scikit-learn library

Table 2: Data Features	
Wave Range	**Wavelength**
Near-Infrared	J(1.22 μm)
	H(1.63 μm)
	K(2.19 μm)
Mid-Infrared	3.4 μm
	3.6 μm
	4.5 μm
	4.6 μm
	5.8 μm^*
	8 μm
	9 μm
	12 μm^*
Far-Infrared	18 μm
	22 μm
	25 μm
	60 μm
	65 μm
	90 μm
	100 μm
	140 μm
	160 μm

* We had two different filters at 5.8 μm and 12 μm in our features from different surveys.

2.3 Shuffling

So far, our data was sorted by their labels. To prepare them for training, we shuffled them to ensure that each data point would create an independent change on the model, without being biased by the same points before them.

2.4 Scaling

The next step in data pre-processing was to scale the features so that values from different IR surveys would not impact the model in case they have different calibrations. We used Robust Scaler, Standard Scaler, and MinMax Scaler for this purpose.[3]

[3]From scikit-learn library

2.5 PCA

As we said, our data contains 22 features. To extract information from such a high-dimensional space, we can project it into a lower-dimensional sub-space by applying PCA on it. It keeps essential parts of data that have more variation and removes non-essential parts that have fewer variation. As our last step in data per-processing, we used PCA [4] method so that it became better prepared for training. In section 3.2 when we draw the scatter matrixes, we will see how applying PCA improves the quality of data categorization.

So far we came to a tidy dataset of about 77900 IR sources each having 22 features. You can see a summary of our final data and its initial sources in Table 3.

Table 3: Data Sources

Label	**Number of Samples**	**Source**	**Format**
YSO	56000	VizieR	.tsv
O-rich AGB	16000	Suh & Kwon	.csv
		IRSA	.html
		ESA	.fits
		SAGE	.dat
C-rich AGB	5600	Suh & Kwon	.csv
		IRSA	.html
		ESA	.fits
		SAGE	.dat
S-rich AGB	300	Suh & Kwon	.csv
		IRSA	.html
		ESA	.fits
		SAGE	.dat
post-AGB	1700	VizieR	.tar.gz
Total	**77900**		**.csv**

Our goal in the next section would be to get to know the properties of the data by visualization.

[4] From scikit-learn library

3 Statistical Analysis

3.1 Scatter Matrix

To get a sense of how features are related to each other and how they change with respect to one another, we can look at their histograms and also the scatter plots of each feature versus the other ones. Both tasks can be done by plotting a Scatter Matrix [5]. To see wether it makes a difference or not, we plotted scatter matrix once before scaling, once after scaling, and once after PCA. The results are shown in Figure 1, Figure 2, and Figure 3.

3.2 Correlation Matrix

After having "seen" the data, we plotted the Correlation Matrix [6] to have a numerical measure of how correlated our features are and to understand the relationships between features. Again, we executed this before scaling, after scaling, and after PCA. The results are shown in Figure 4, Figure 5, and Figure 6.

4 Relevant Literature

Getting to know the infrared sky is an important step toward understanding many astrophysical phenomena —such as star formation and late stages of stellar evolution. Many infrared sources such as protostars, evolved dusty stars, and young stars have not yet been fully characterized. If we want to increase the potential for extracting scientific data from IR surveys, we have to gain more information about infrared sources.

Numerous articles in astrophysics have somehow dealt with the identification or classification of infrared sources. In the last section, we refer to a few of them that are kind of related to our work.

[5] From scikit-learn library

[6] From scikit-learn library

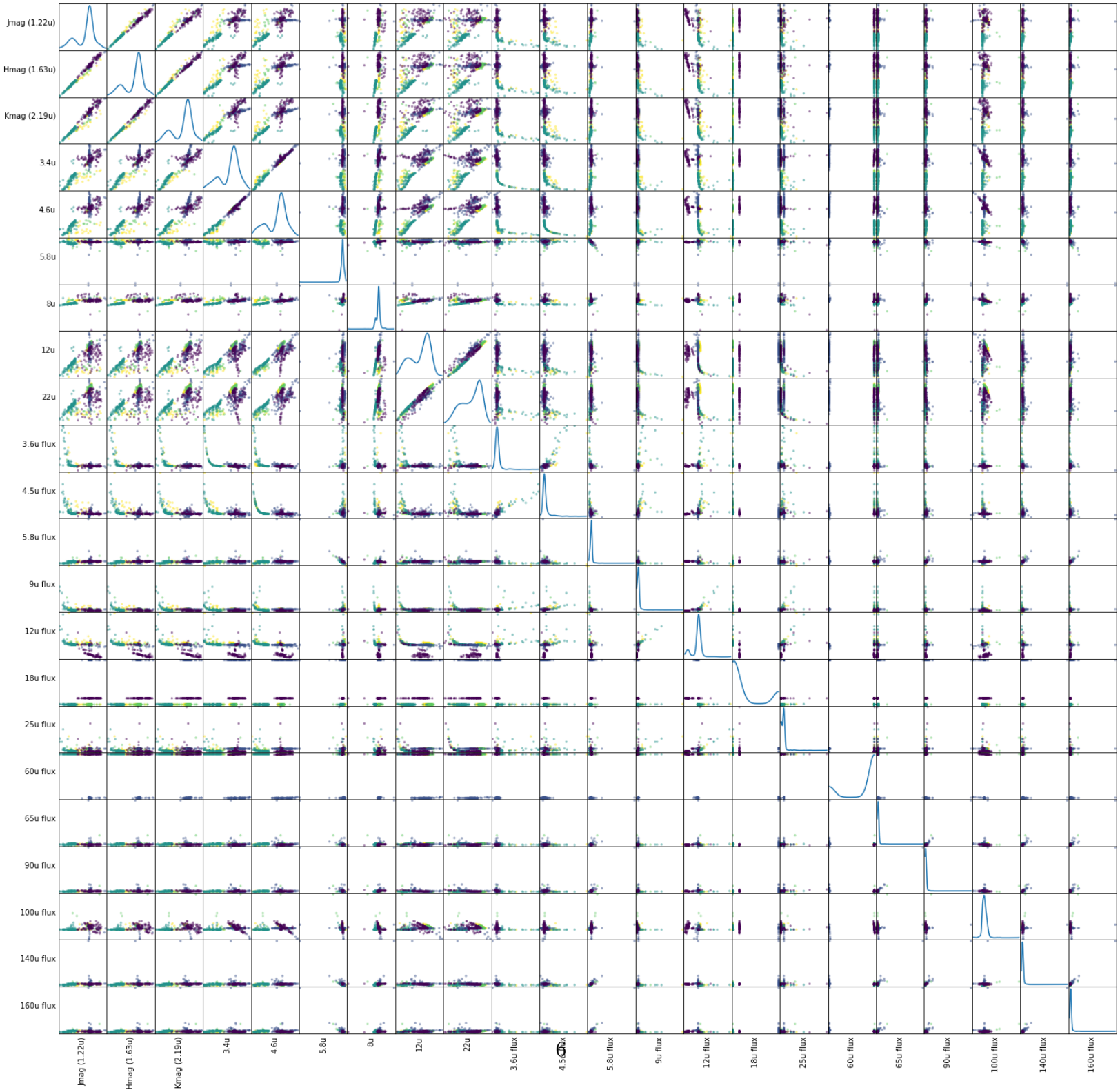

Figure 1: The scatter matrix of data *before* standard scaling. Different colors refer to different labels.

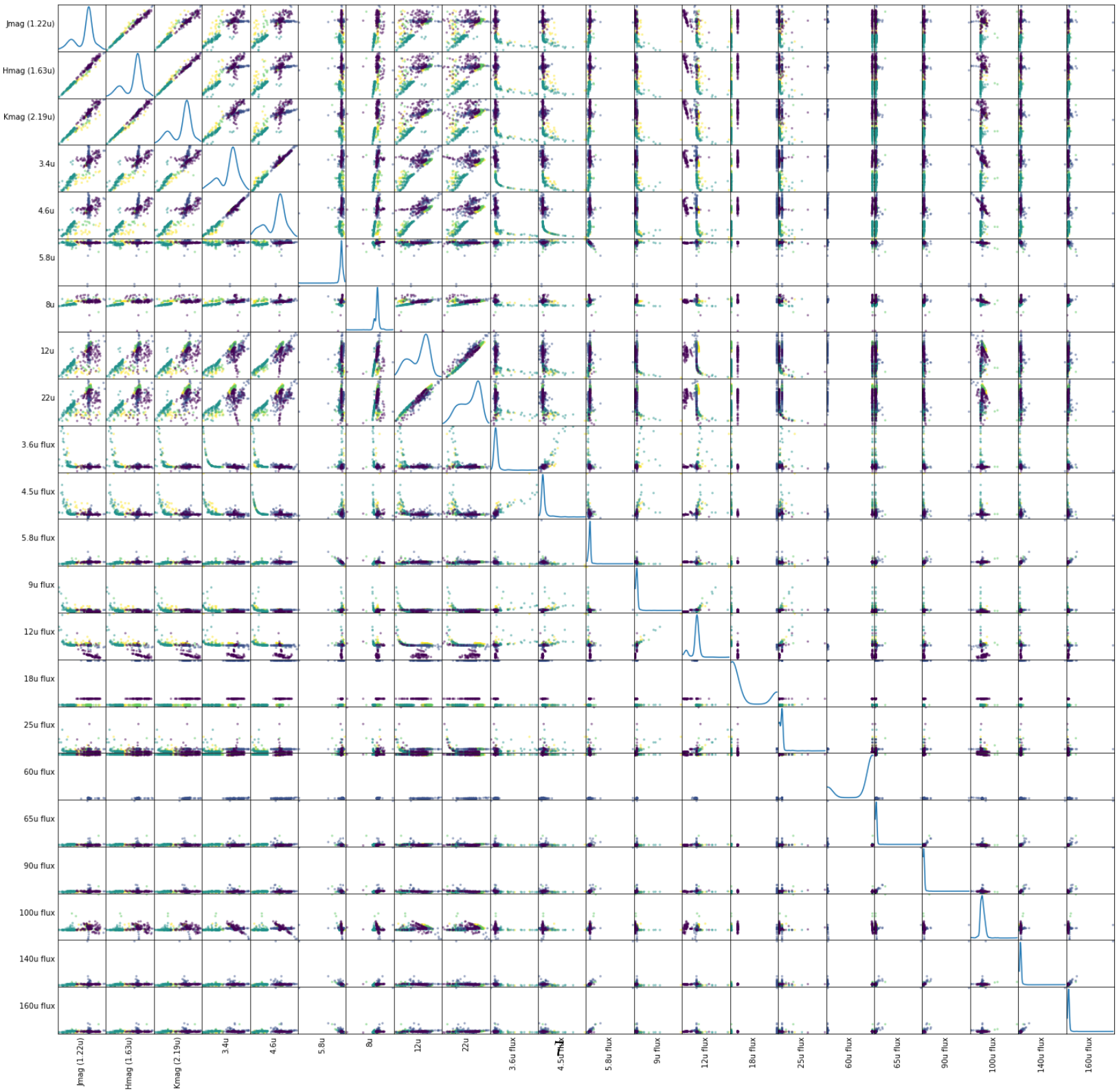

Figure 2: The scatter matrix of data *after* standard scaling. Different colors refer to different labels.

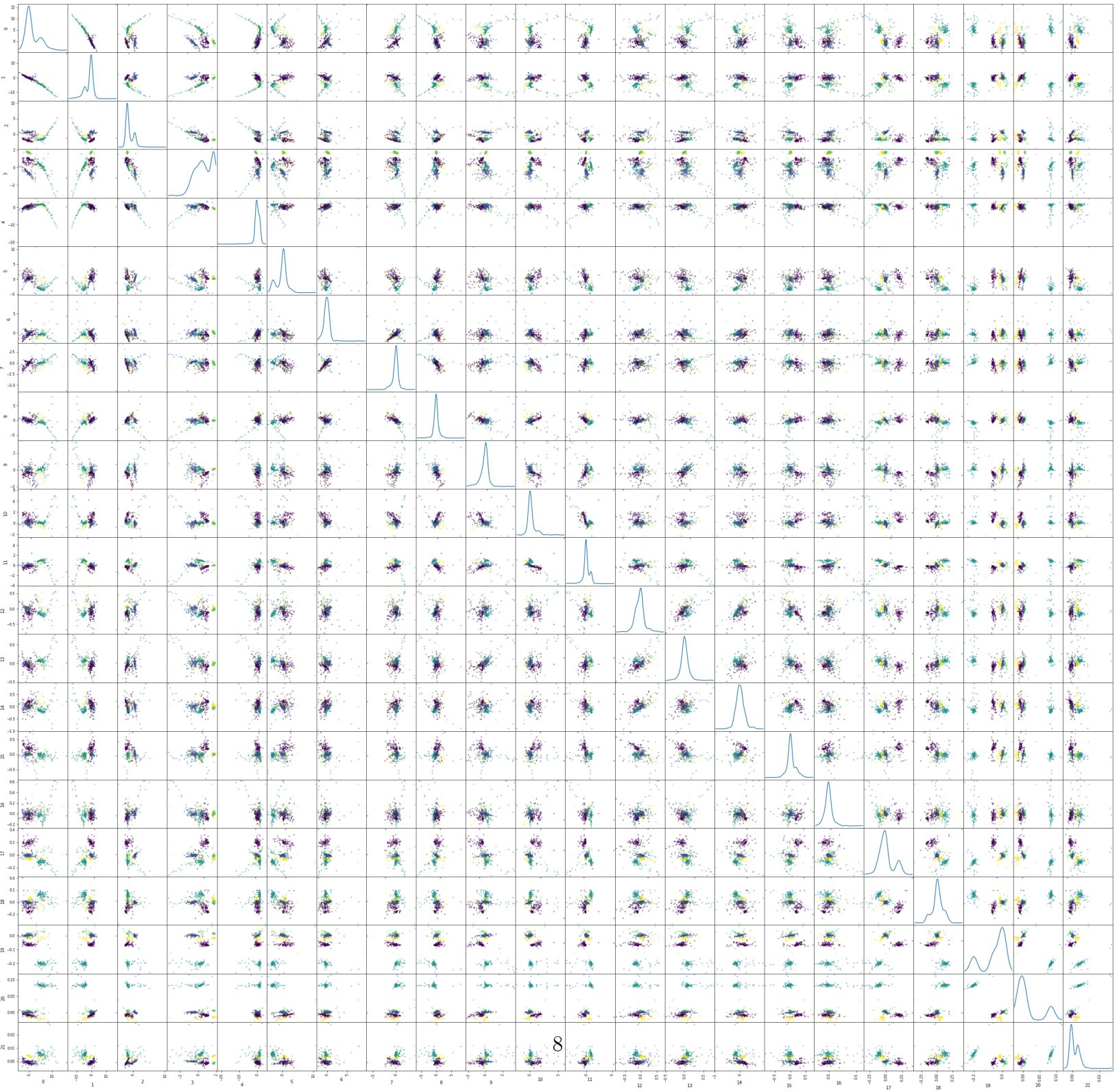

Figure 3: The scatter matrix of data *after* PCA. Different colors refer to different labels.

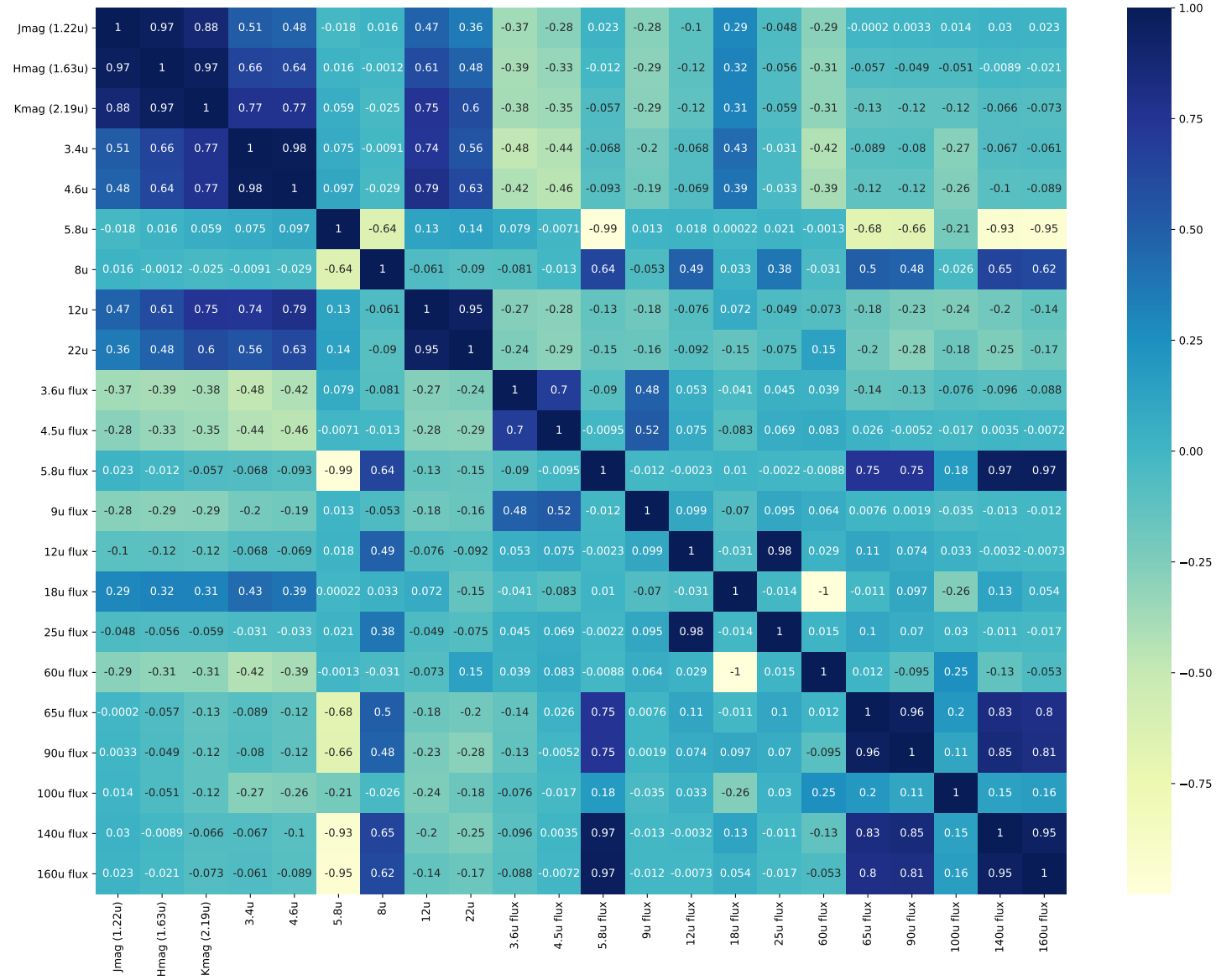

Figure 4: The correlation matrix of data *before* standard scaling. Different colors refer to different labels.

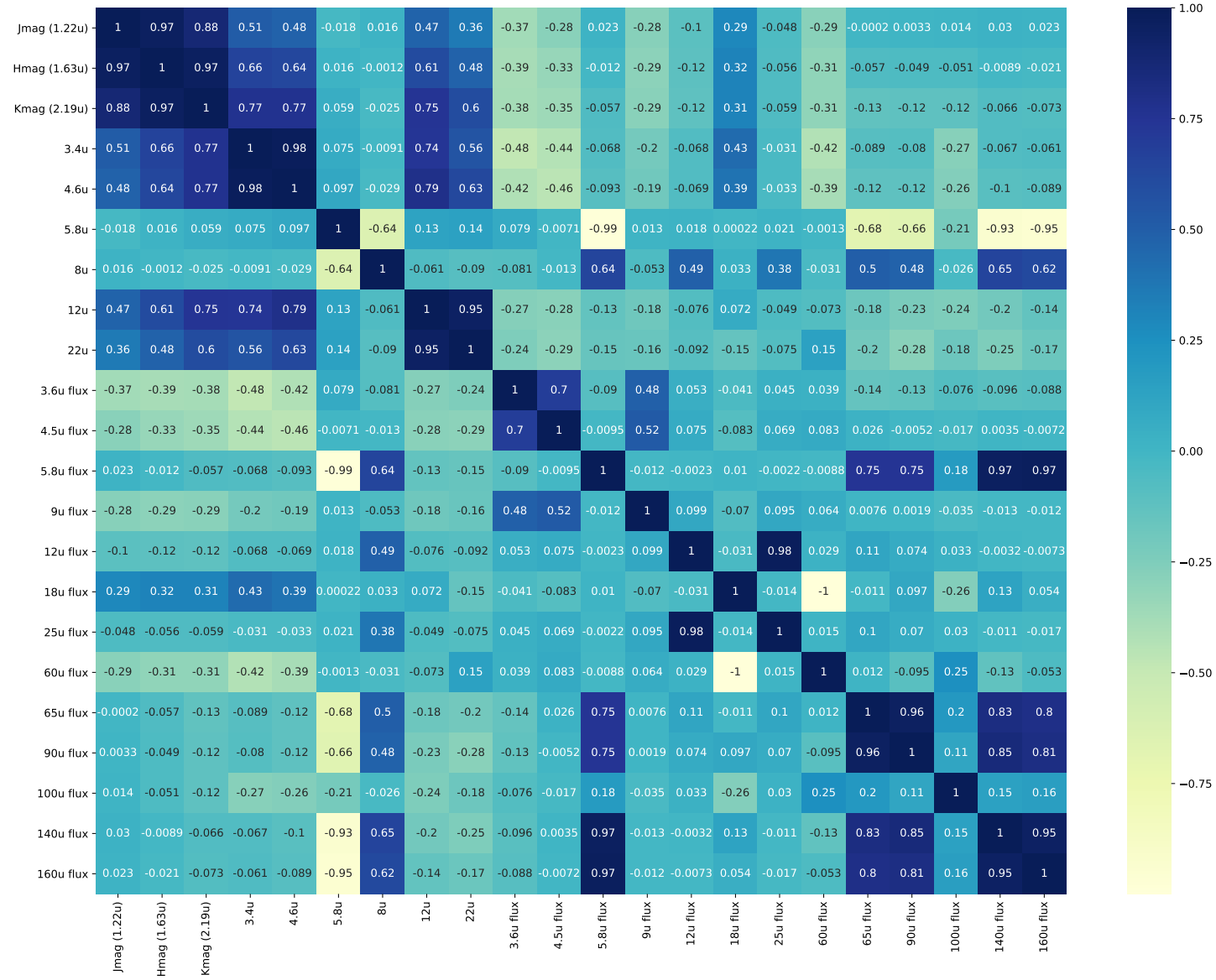

Figure 5: The correlation matrix of data *after* standard scaling. Different colors refer to different labels.

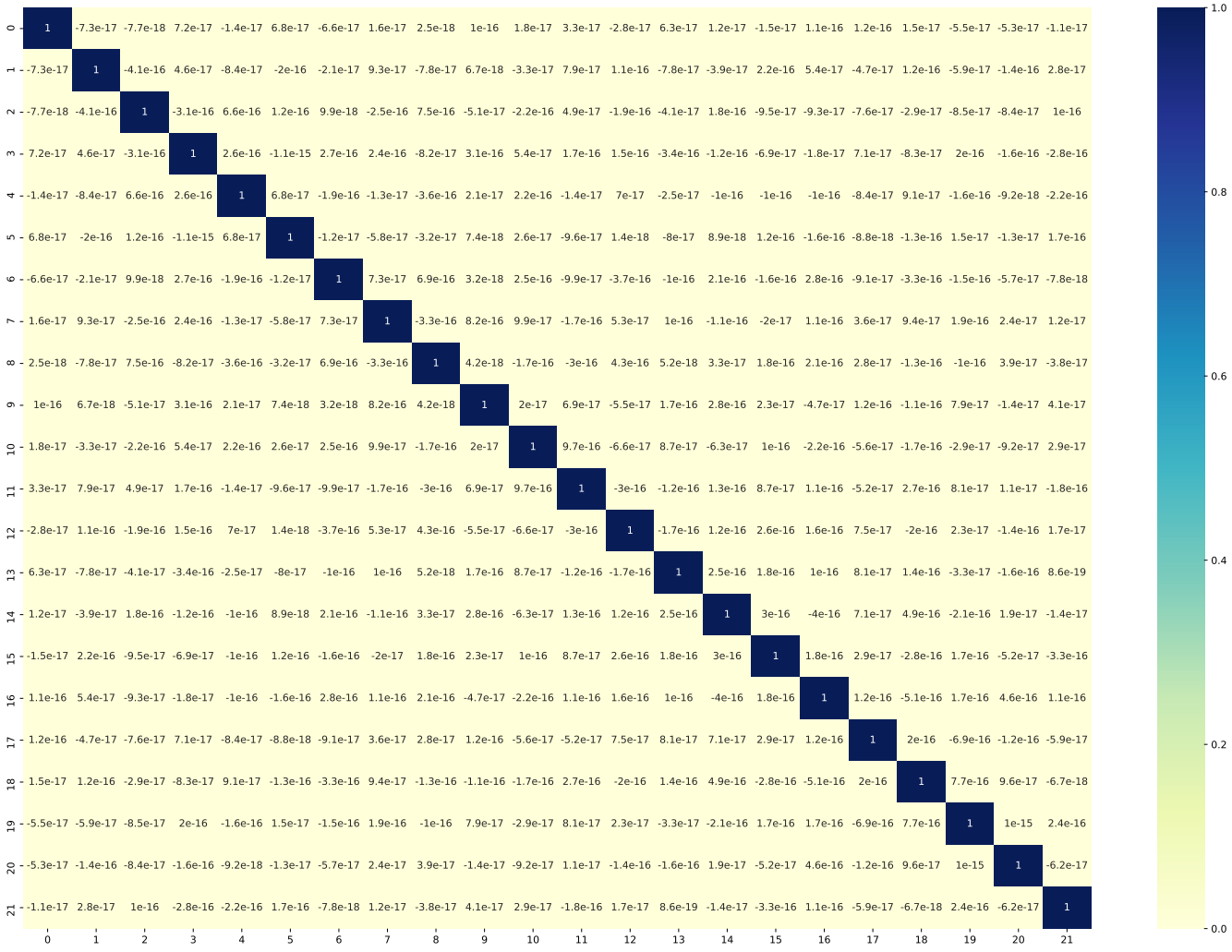

Figure 6: The correlation matrix of data *after* PCA. Different colors refer to different labels.

4.1 Dust Disks Around YSOs

Young stellar objects are surrounded by a gas and dust disk as a consequence of the star formation process. With the help of astrophysical theoretical models such as radiative transfer model it is possible to perform calculations to reproduce these dust disks. Indeed their spectral energy distribution (SED) is investigated. Such studies require various infrared observational data from IR surveys. Kyung-Won Suh (2008) and Kyung-Won Suh (2016) are samples of such studies. In this literature you would see the term TTS being repeated. It stands for "T Tauri Stars" that are generally known to be low-mass pre-main sequence (PMS) stars.

4.2 Dusty Stellar Winds from AGBs

An asymptotic giant branch (AGB) star is the last evolutionary phases of a low mass star. The cool envelopes around AGB stars are believed to be the main site of dust formation. Nearly all AGB stars can be identified as long-period variables (LPVs) with large amplitude pulsation. Shock waves produce by the strong pulsation and radiation pressure on newly formed dust grains drive dusty stellar winds with high mass-loss rates. IR observations of AGB stars identify various dust species in different physical conditions. Kyung-Won Suh (2014) reviews the complete astrophysiscs of AGBs including their classifications.

4.3 Machine Learning in Astronomy

Astronomical datasets have been growing at an exponential rate. These huge amounts of data need analysis: pattern recognition, prediction, classification, and much more. Classification problems can be solved by Machine-Learning models. Machine-learning has the ability to increase the explanatory power of photometric observations, which may be less informative about each individual object, but provides more data on more objects in general. Baron (2019) is an review articles to cover basic topics in supervised machine learning.