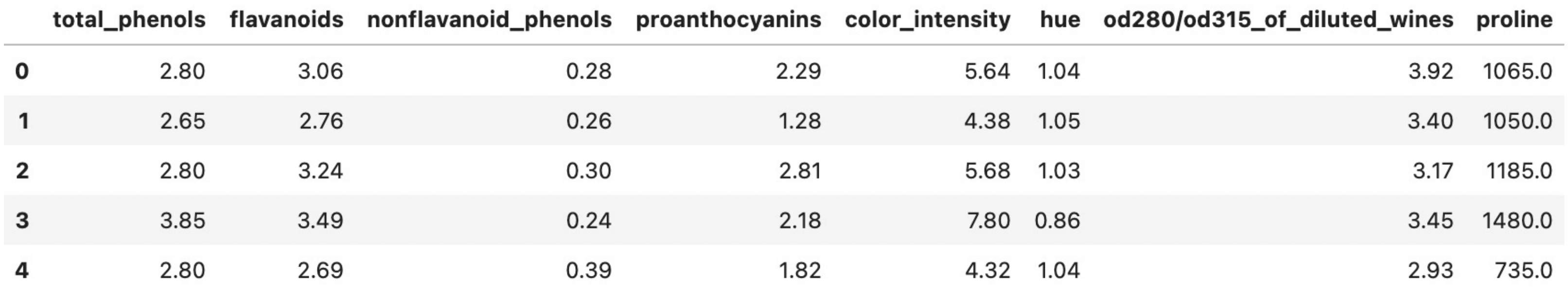
Scaling

Data Science & Analytics
University of Missouri

Feature Scaling

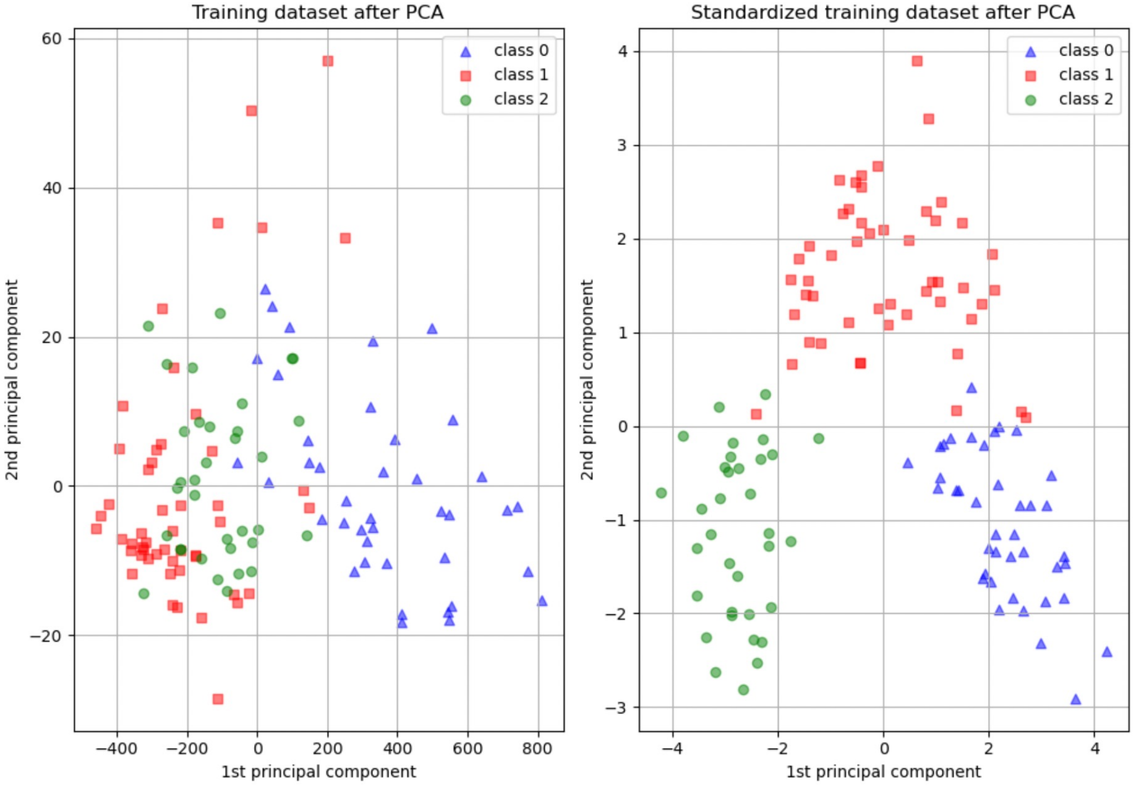
- **An Important Step in ML**
- **Brings all the features in the same scale/range**
  - E.g., consider 3 features in a dataset: age, num of pets, yearly salary
    - Increase in age by a year != increase in salary by $1

Importance of Feature Scaling



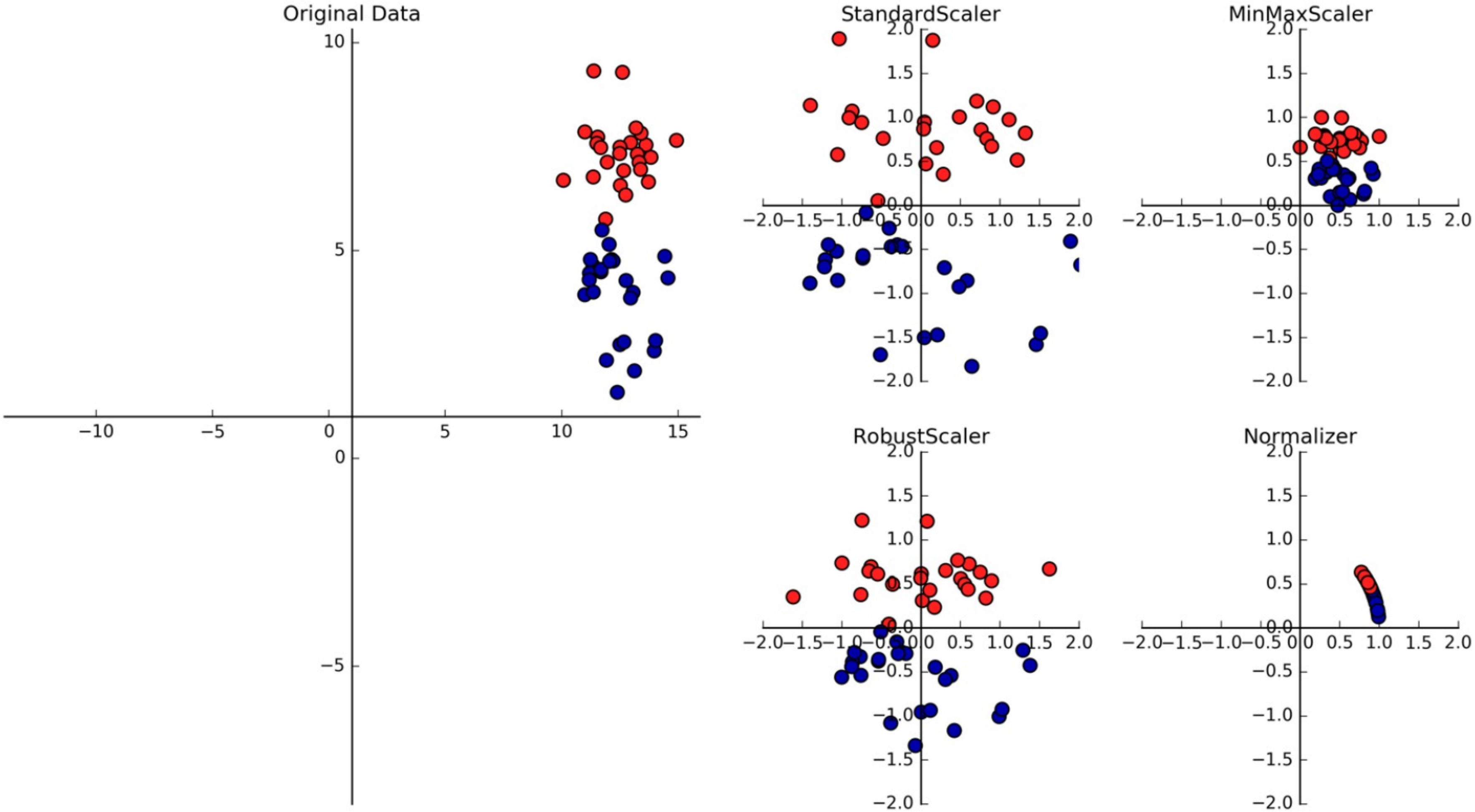
	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline
0	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065.0
1	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050.0
2	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185.0
3	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480.0
4	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735.0

Acc = 81%



Acc = 98%

Types of Scaling



Standard Scalar

- Centers the data by using the following formula, where *u* is the mean and *s* is the standard deviation

$$x\_scaled = (x - u) / s$$

MinMax Scaler

- **Transforms features by scaling each feature to a given range**
  - This range can be set by specifying the *feature_range* parameter
    - default at *(0,1)*
- **Works better for cases where the distribution is not Gaussian or the standard deviation is very small**
- **sensitive to outliers**

x_scaled = (x-min(x)) / (max(x)–min(x))

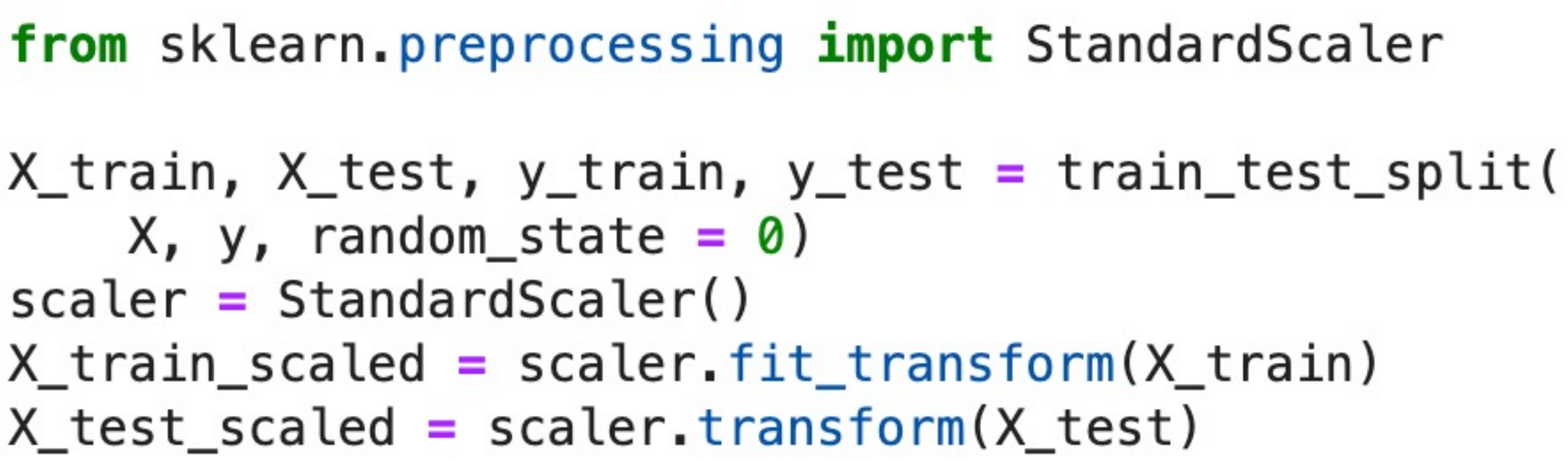
Robust Scaler

- If your data contains many outliers, scaling using the mean and standard deviation of the data is likely to not work very well
- It removes the median and scales the data according to the quantile range.

Normalization

- The process of scaling individual samples to have unit norm
- You need to normalize data when the algorithm predicts based on the weighted relationships formed between data points
- One of the key differences between scaling (e.g. standardizing) and normalizing, is that normalizing is a row-wise operation, while scaling is a column-wise operation.

Sample Code: Scaling



```
from sklearn.preprocessing import StandardScaler

X_train, X_test, y_train, y_test = train_test_split(
    X, y, random_state = 0)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Q. Why is test data not used in the fitting?