

```
In [1]: import numpy as np
import pandas as pd
import random
random.seed(42)
np.random.seed(42)
```

```
In [2]: df = pd.read_csv("wine-clustering.csv")

df.head()
```

Out[2]:

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflav
<b>0</b>	14.23	1.71	2.43	15.6	127	2.80	3.06	
<b>1</b>	13.20	1.78	2.14	11.2	100	2.65	2.76	
<b>2</b>	13.16	2.36	2.67	18.6	101	2.80	3.24	
<b>3</b>	14.37	1.95	2.50	16.8	113	3.85	3.49	
<b>4</b>	13.24	2.59	2.87	21.0	118	2.80	2.69	



```
In [4]: X = df.copy()
numerical_cols = X.columns.tolist()
```

```
In [5]: from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import LocalOutlierFactor
```

```
In [6]: num_pipe = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler())
])

processor = ColumnTransformer(transformers=[
    ("num", num_pipe, numerical_cols)
], remainder="drop")

lof_pipe = Pipeline(steps=[
    ("process", processor),
    ("lof", LocalOutlierFactor(n_neighbors=20, contamination=0.05))
])

labels = lof_pipe.fit_predict(X)

df["outlier"] = labels
```

```
In [7]: df["outlier"].value_counts()
```

```
Out[7]: outlier
      1    169
     -1     9
Name: count, dtype: int64
```

```
In [8]: submission = pd.DataFrame({
        "Id": df.index,
        "outlier": df["outlier"]
    })

submission.to_csv("submission.csv", index=False)
```