(発展) 課題 2

(Adv) cs3-06-assign2

データ winequality-red.csv は、ワインの品質に関するデータである。これを用いて、以下の操作を行うスクリプトをJupyterで作成せよ(ノートブック名は cs3-06-assign2.ipynb/html とせよ)

1. 必要なライブラリをimportする。
2. winequality-red.csv のデータをデータフレーム df に読み込み、行数と列数、各列のデータ型と欠損値でないデータの数、先頭5行を表示して確認。
3. df の各列の統計量を表示 (describe()を用いる)。
4. df の各列を平均0, 母標準偏差(偏差二乗和をデータ数Nで割った分散の平方根)1に標準化し、変数 X_scaled に代入。
5. 手順5 で得た X_scaled の平均と母標準偏差を表示。
6. X_scaledに、df と同じ列ラベルを付与したデータフレーム df_scaled を作成し、先頭5行を表示。

(Adv) cs3-06-assign2

7. df_scaledに対して、KMeans法 (n_init=10とする) によるElbow法を、最大クラスタ数10で実施。クラスタ数を横軸、Inertiaを縦軸とするグラフを描画。
8. データフレーム df_scaled に対してクラスタ数4で KMeans法によるクラスタリングを実行。このとき、n_init=10, random_state=5とせよ。
9. 元のデータフレームdfに、新たな列 cluster_no を追加し、各データが属するクラスタの番号を格納する。先頭5行を表示して確認。
10. 「cluster_no」列の各値の出現数(各クラスタのメンバー数) を表示。
11. 「cluster_no」列以外の列について、クラスタごとの分布を箱ひげ図で確認。クラスタ番号0 が他のクラスタと区別されるのに大きく寄与していると思われる列を2つ挙げよ。