データサイエンス特論　授業課題　第五回分（クラスタリング）

授業課題１で紹介した IRIS データに対して、K-means クラスタリング手法を適用し、データ内の target 値を真値とし、クラスタリングの評価尺度の１つである Purity 値を計算せよ。クラスタリングの実行にあたっては、Python 言語の適当なパッケージにある関数を使ってよいとする。また生成するクラスター数は実際の IRIS の種類に合わせて３とすること。締切は、（祝日であるが）一週間後【7 月 18 日(月)】の深夜(23:59)までで、Moodle LMS にアップロードすること。今回は、Python コード、実行結果（150 個のデータがそれぞれ何番のクラスターに割り当てられたか）と Purity 値を合わせ、提出すること。Jupyter notebook(拡張子.ipynb も可、ただし実行結果や Purity 値が notebook を起動すると見えること（すなわち、こちらが実行しなくても確認できること）を条件とする)。

なお、一般的には、K-means 法は、実行するたびに結果が少し変わることがあるので、何回か実行した結果の平均の Purity 値でも OK とする。（IRIS データでは、ほぼ変わらないと察します）

『コメントとヒント』: 今回同時にプログラム課題１でクラスタリングを課しています（プログラム課題のほうは、締め切りは２週間)。そちらの資料を適宜、参照ください。プログラム課題と合わせ、推奨する作業手順は、まず、こちらの（数値だけの）IRIS データ（３クラス（カテゴリ））の小規模データでクラスタリングを行ってから、プログラム課題にある 20newsgroup（テキストデータ）にチャレンジするのがいいかと思います。