**Grupowanie** - jedna z metod eksploracji danych, oznacza grupowanie rekordów, obserwacji lub przypadków w klasy podobnych obiektów. Celem metod grupowania jest identyfikacja grup rekordów aby ich podobieństwo było bardzo duże wewnątrz grupy, a do rekordów z innej grupy bardzo małe. Chodzi o to by stworzyć grupy, dla których zmienność pomiędzy grupami będzie duża w porównaniu do zmienności wewnątrzgrupowej.

**Grupa** - to zbiór rekordów podobnych do siebie nawzajem i niepodobnych do rekordów z innych grup.

**Różnica między klasyfikacja a grupowaniem**

Brak zmiennej celu w przypadku grupowania. Głównym zadaniem algorytmu grupowania jest dokonanie podziału zbioru na stosunkowo zgodne grupy.

**Analiza skupień** - to technika pozwalająca zredukować przestrzeń przeszukiwań dla algorytmów.

Aby przeprowadzić analizę skupień należy wziąć pod uwagę:

1. Jak mierzyć podobieństwo.
2. Jak kodować zmienne jakościowe.
3. Jak normalizować zmienne ilościowe.
4. Ile grup należy się spodziewać

**Podział algorytmów grupowania**

* hierarchiczne
* niehierarchiczne

**Grupowanie hierarchiczne** - polega na tworzeniu drzewiastej struktury poprzez rekurencje:

* łączenie (**metody aglomeracyjne**)
* dzielenie (**metody rozdzielające**)

istniejących grup.

**Metoda aglomeracyjna** - (ześrodkowanie) zakłada, że początkowo każda obserwacja to osobna grupa składająca się z jednego elementu. W kolejnych krokach dwie grupy będące najbliżej siebie łączy się w nową wspólną grupę. W każdym kroku uzyskuje się redukcję (o jeden) liczby skupień (grup) w zbiorze danych.

**Metody rozdzielające** - cały zbiór umieszczany jest w jednej grupie. W kolejnych krokach rozdziela się rekordy najbardziej niepodobne do siebie na oddzielne dwie grupy. Algorytm kończy działanie do momentu kiedy powstaną grupy jedno elementowe.

**Kluczowa rola to mechanizm określania odległości między grupami rekordów.**

**Określanie odległości**

**Metod pojedynczego połączenia** - metoda najbliższego sąsiedztwa. Określa odległość pomiędzy najbliższymi punktami. Oparta na minimalnej odległości pomiędzy dowolnym rekordem grupy A i dowolnym rekordem z gr. B. (Tendencja do tworzenia długich cienkich grup).

**Metoda całkowitego połączenia** - metoda najdalszego sąsiedztwa. Określa ona odległości pomiędzy najbardziej oddalonymi punktami. Oparta na maksymalnej odległości pomiędzy dowolnym rekordem z grupy A i dowolnym rekordem z gr. B. (Tendencja do tworzenia zwięzłych kulistych grup).

**Metoda średniego połączenia** - oparta na wyliczaniu średniej odległości wszystkich rekordów z grupy A do wszystkich rekordów z gr. B. Metoda pozwala na ograniczenie wpływu rekordów o wartościach ekstremalnych na końcowy wynik grupowania.

**Algorytm *k*-średnich**

1. Ustalenie na ile grup (*k*) należy podzielić zbiór danych.
2. Losowy wybór *k* rekordów z przeznaczeniem na środki grup.
3. Znalezienie najbliższego środka grupy dla każdego rekordu.
4. Znalezienie (dla każdej z *k* grup) centroidu grupy i uaktualnienie położenia każdego środka grupy tak, aby stała się nim wartość znalezionego centroidu.
5. Powtarzanie kroków 3 i 4 do zbieżności lub planowanego zakończenia algorytmu.

**Centorid** - środek ciężkości wszystkich punktów w grupie.