Moduł 7
Narzędzia do analizy Big Data, część 2: Przegląd analityki Big Data

iBigWorld:
Innovations for Big Data in a Real World

ULSIT team

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Określenie celów

Wykład stanowi podstawowe, ale wyczerpujące wprowadzenie do:

- analiza danych poprzez statystyki podsumowujące i metody wizualizacji
- Terminologia analizy danych
- Metody uczenia maszynowego i ich związek z zaawansowaną analityką

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Struktura wykładu

- Terminologia analizy danych
- Analiza danych (EDA)
- Analiza danych poprzez statystyki podsumowujące
- Analiza danych za pomocą wykresów i wykresów
- Jakość danych
- Podejścia, techniki i algorytmy

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wyniki nauczania

Umiejętności:

- Opisz analizę danych
- Opisz rolę statystyk podsumowujących w analizie danych
- Opisz rolę wizualizacji w analizie danych
- Wyjaśnij, jakie działania są podstawą przygotowania danych
- Wyjaśnij, czym jest jakość danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Terminologia analityki danych

- Próbki i zmienne

Variables (features/attributes/columns/fields/dimensions)

Samples (records/examples/rows/instances/observations)

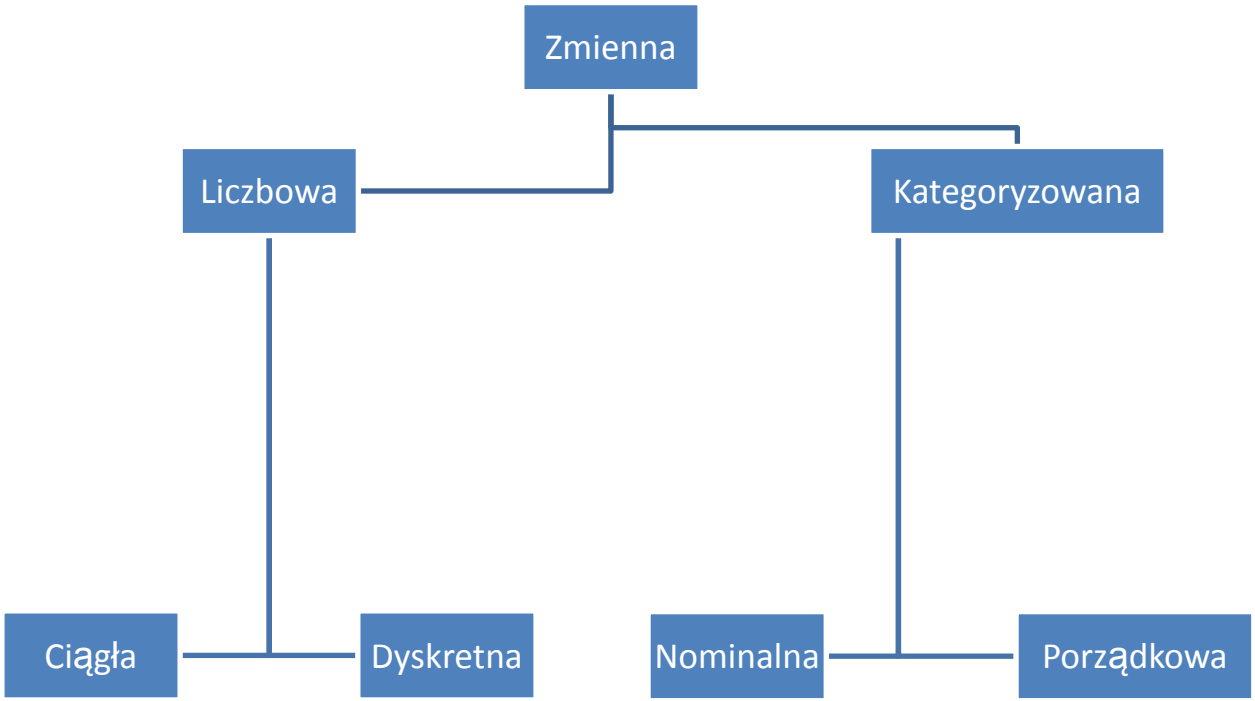
ID	Sex	Age	Educational Degree	Annual Income
354	M	33	Master	55,000
123	F	25	Diploma	23,000
202	F	48	Diploma	34,000
304	M	55	Bachelor	41,000

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Terminologia analityki danych

- Typy zmiennych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Terminologia analityki danych

- Właściwości wag pomiarowych

Property	Nominal	Ordinal	Interval	Ratio
Named	✔	✔	✔	✔
‘Natural’ order		✔	✔	✔
The mode can be measured	✔	✔	✔	✔
Median can be measured		✔	✔	✔
Mean can be measured			✔	✔
The exact difference between values			✔	✔
“True zero” value				✔

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza danych

Kategorie

- **Korelacja** dostarcza informacji o relacji między zmiennymi w danych
- **Trendy** pokazują charakter relacji w danych
- **Wartości odstające** to punkty, które są znacznie mniejsze lub większe niż pozostałe dane. Mogą znacząco zmienić rodzaj rozkładu, a tym samym wpłynąć na wyniki analizy. Czasami jednak przedmiotem badań są same emisje.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza danych

Metody

- **Statystyki podsumowujące** dostarczają ważnych informacji o średnich, głównych trendach i rozproszeniu.
- **Wizualizacja** danych zapewnia wizualną reprezentację danych.

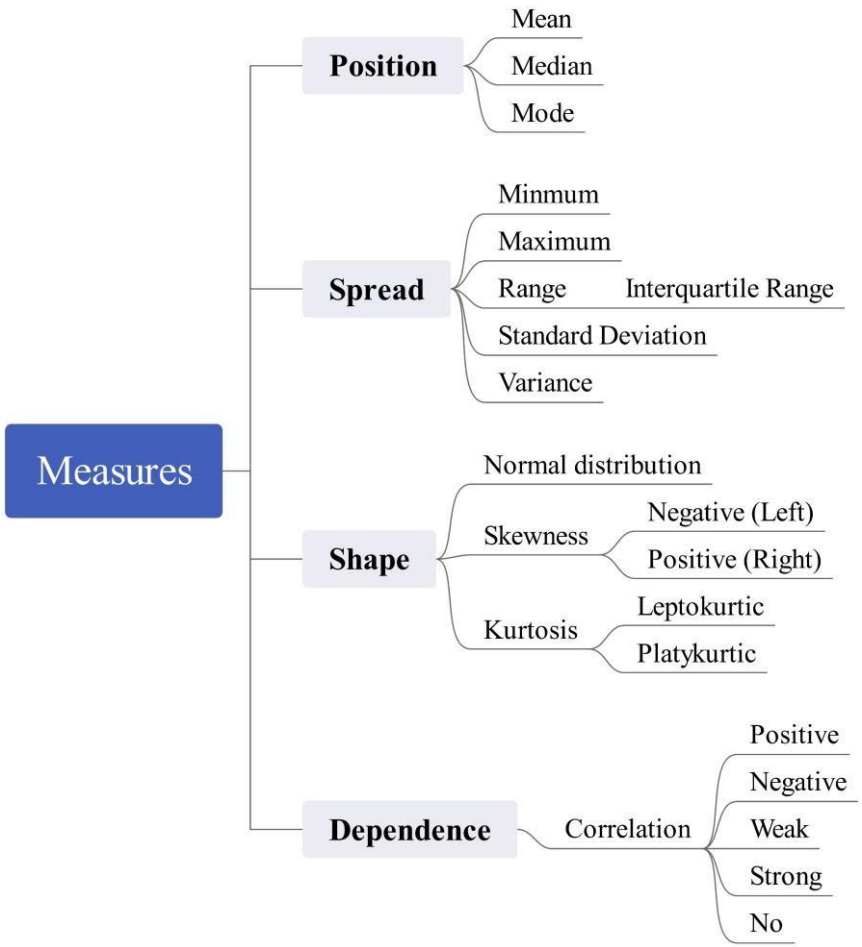
Statystyki podsumowujące i wizualizacja powinny być używane razem do analizy zbioru danych.

Z kolei lepsze zrozumienie kieruje resztą procesu i prowadzi do bardziej pouczającej analizy.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

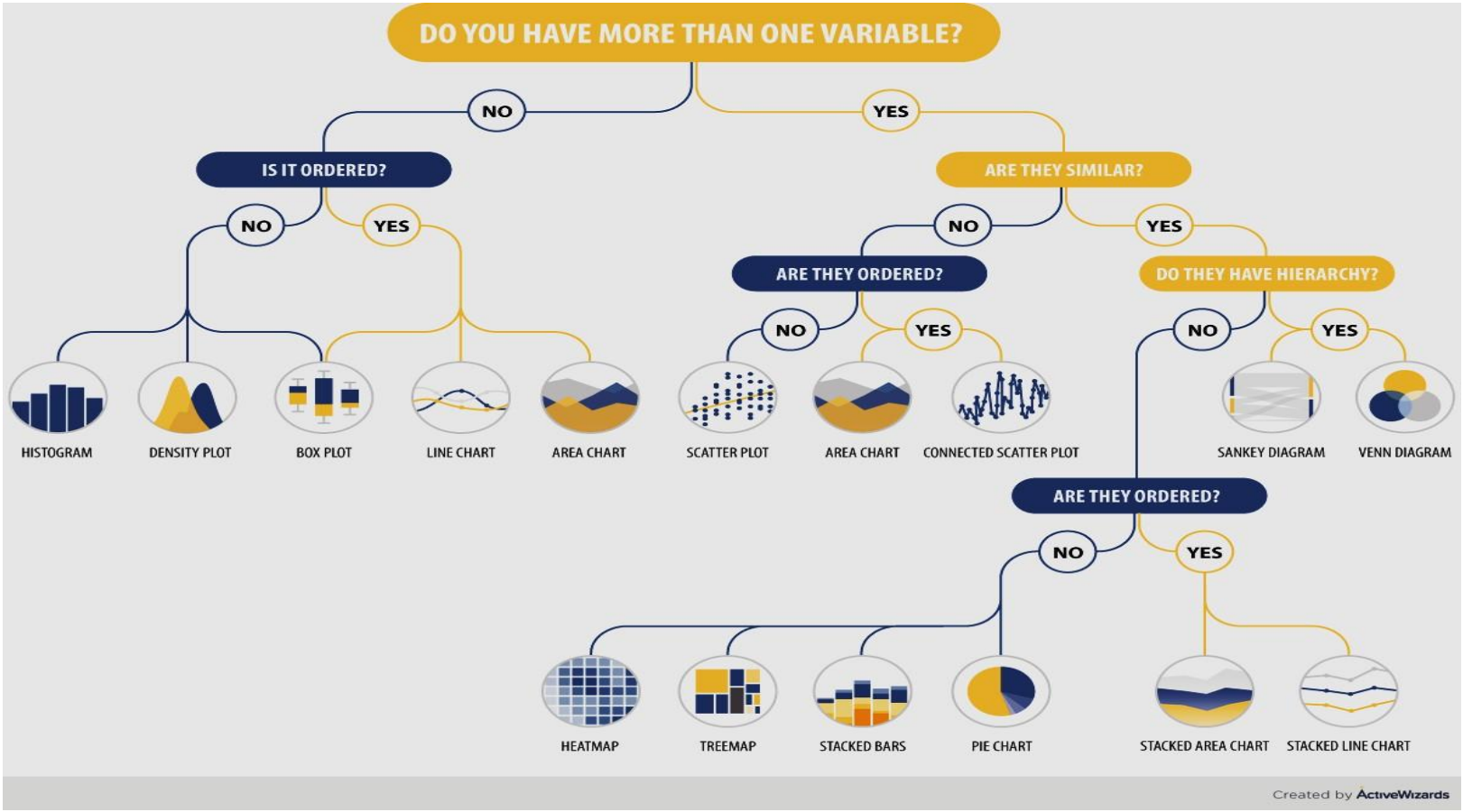
Statystyki podsumowujące w analizie danych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

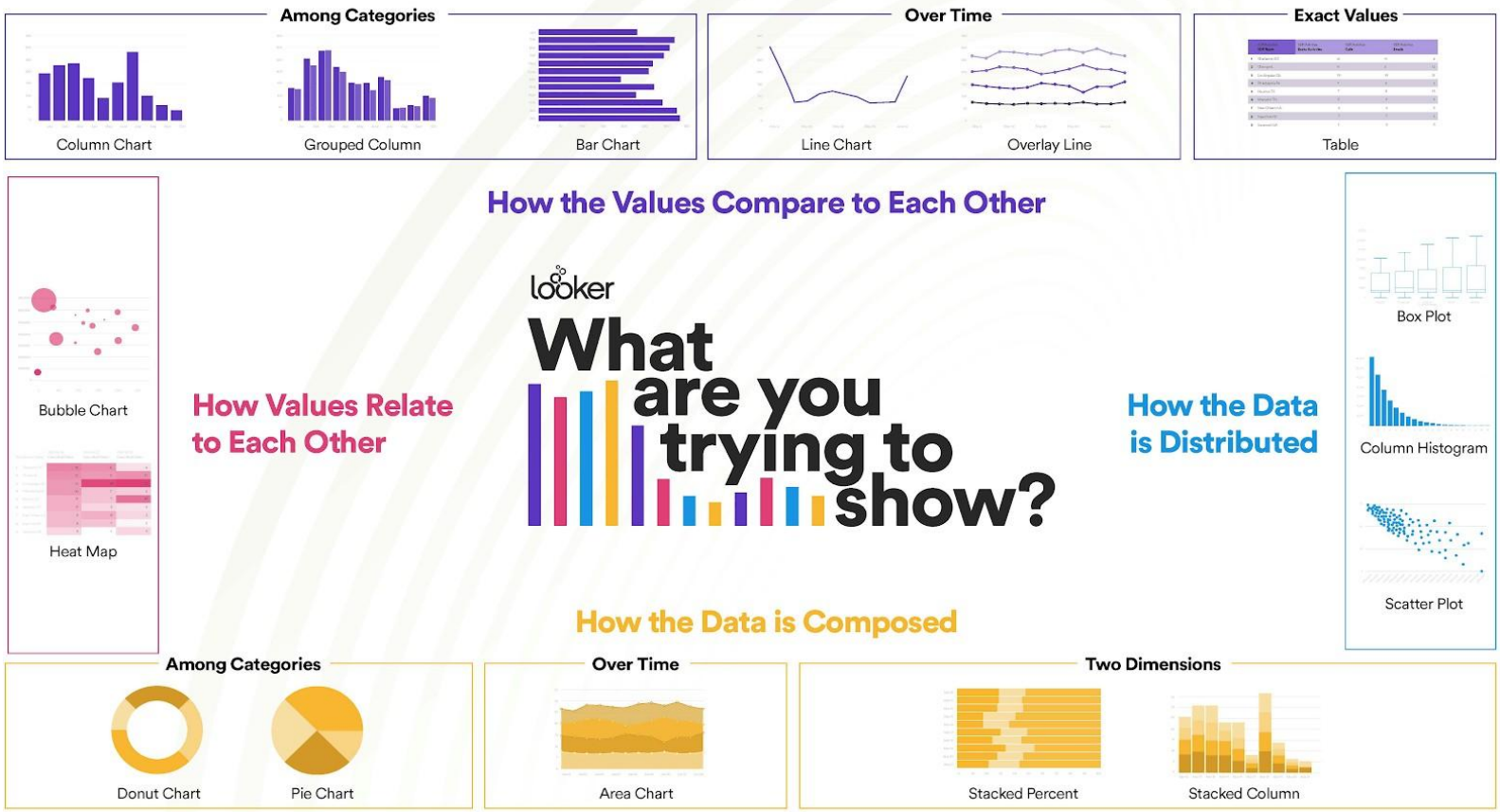
Analiza danych w języku wykresów i diagramów



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Wykresy do analizy danych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza danych za pomocą wykresów i wykresów:

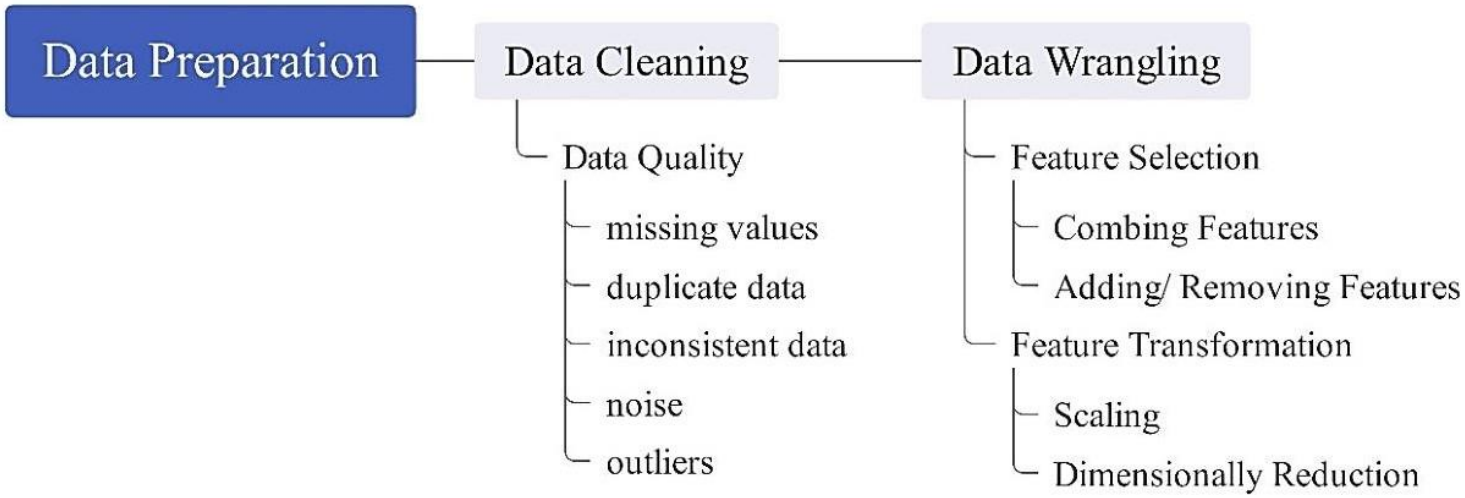
- Wizualizacja danych odbywa się w celu dostarczenia użytkownikowi końcowemu. Powinien przedstawiać dokładny, kompletny i prosty obraz poparty danymi, przy jednoczesnym zachowaniu dostępności zrozumienia i odwołania. Istnieje wiele metod konstruowania wykresów i wykresów do wizualizacji danych. Wybór odpowiednich narzędzi wymaga dobrego zrozumienia Twojej firmy, zbioru danych i potrzeb.
- Wizualizacja danych to nauka i sztuka ułatwiania zrozumienia danych i wyciągania z nich wniosków.
- Idealna wizualizacja zapewnia odpowiednią ilość danych, we właściwej kolejności i we właściwej formie wizualnej, aby wyróżnić najważniejsze informacje.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jakość danych

- Źródła zapewnienia jakości danych (działania)



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Jakość danych

- Wiedza o domenie jest niezbędna do podejmowania świadomych decyzji o tym, jak najlepiej przypisywać brakujące wartości, jak obsługiwać zduplikowane rekordy i nieprawidłowe dane oraz co zrobić z szumem i wartościami odstającymi w danych.

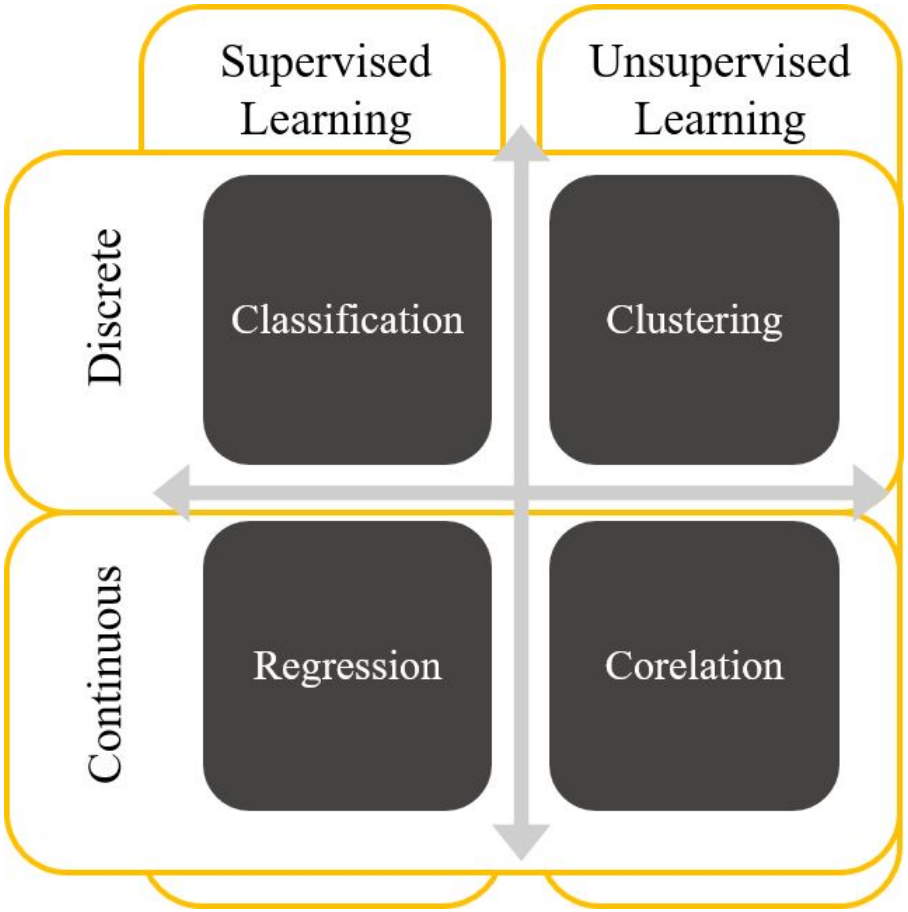
- Pamiętaj:
  - Śmieci wewnątrz = śmieci na zewnątrz!!!
- Aby zapewnić jakość danych wejściowych, należy poświęcić wystarczająco dużo czasu, jest to gwarancja jakości wyników analizy.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Uczenie maszynowe do analizy Big Data

- Podejścia



iBigWorld:2020-1-PL01-KA203-082197

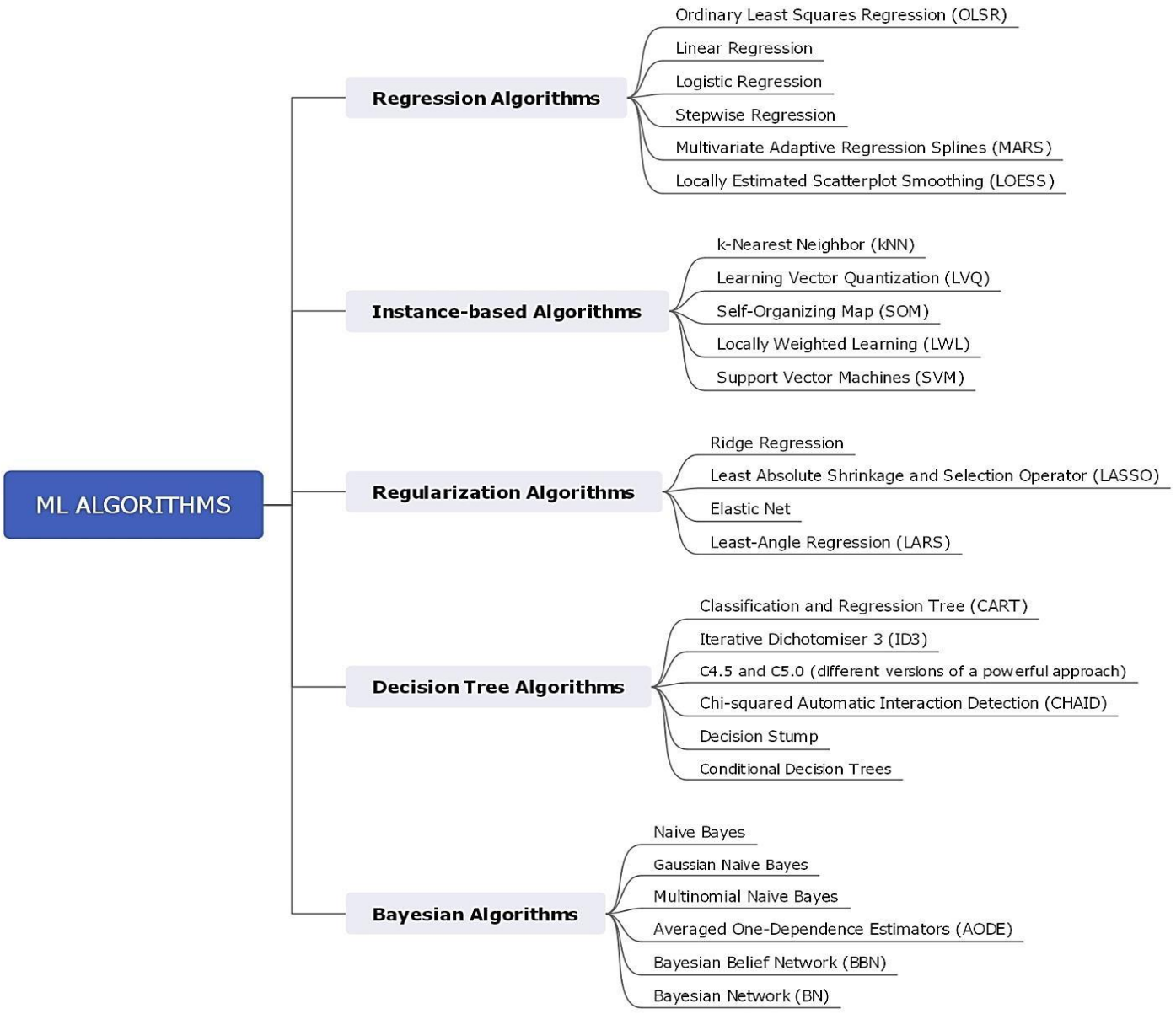
Co-funded by the Erasmus+ Programme of the European Union

Uczenie maszynowe do analizy Big Data

- Na najwyższym poziomie podejścia do uczenia maszynowego dzielą się na dwie klasy: nadzorowane i nienadzorowane.
- W uczeniu się z nauczycielem cel jest zdefiniowany, podczas gdy w uczeniu się bez nauczyciela cel jest nieznany lub niezdefiniowany. Uczenie się z nauczycielem jest zwykle związane z klasyfikacją lub regresją, a uczenie się bez nauczyciela jest zwykle związane z analizą skupień lub asocjacyjną.
- W klasyfikacji obiekt należy do kategorii, a w regresji określana jest wartość liczbowa. Celem grupowania jest podzielenie obiektów na grupy. Analiza asocjacyjna ustala związek między obiektami lub zdarzeniami.

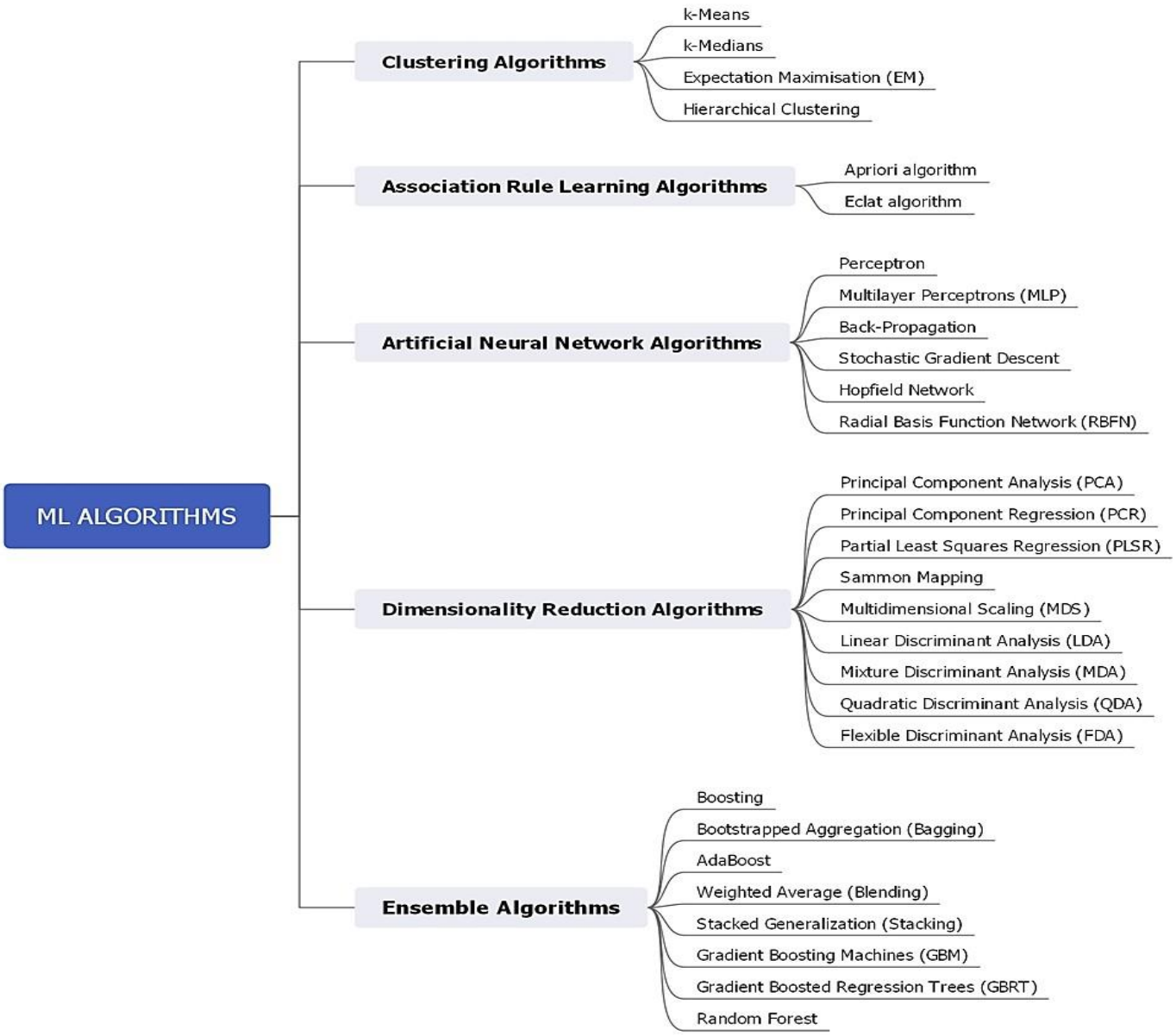
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Uczenie maszynowe do analizy Big Data

- Wskaźniki oceny

Klasyfikacja	**Regresja**
Pewność	*Błąd średniej kwadratowej*
Precyzja	Pierwiastek z bładu średniokwadratowego
Wrażliwość	*R-kwadrat*
F1-wskaźnik	*Średni błąd bezwzględny*
Obszar pod krzywą (AUC)	

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Uczenie maszynowe do analizy Big Data

- Uczenie maszynowe ma na celu: badanie danych w celu zidentyfikowania wzorców lub klasyfikacji, przewidywanie wyników lub działań, identyfikowanie nieznanych wzorców i relacji, wykrywanie anomalii lub nieoczekiwanych zachowań.
- Najbardziej znane zadania uczenia maszynowego to: regresja i klasyfikacja, klastrowanie i redukcja wymiarowości (wybór i selekcja cech)
- Najważniejsze pytanie brzmi, czy zastosowane podejście, technika i algorytm są odpowiednie do rozwiązania danego problemu?

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Literatura

- P. C. Bruce and A. Bruce, Practical statistics for data scientists : 50 essential concepts. O'Reilly, 2020.
- Enterprise Big Data Framework Guide, Enterprise Big Data Professional. 2018, ISBN: ISBN: 978-90-828958-0-3, Available: www.bigdataframework.org.
- Enterprise Big Data Analyst Guide, V1.2, 2021, ISBN:978-90-828958-10.
- W. L. Chang and NBD-PWG, “NIST Big Data Interoperability Framework: Vol 1, Definitions,” 2019, DOI: 10.6028/NIST.SP.1500-1r2.
- Simplilearn, Introduction to Data Analytics, https://www.simplilearn.com/.
- UC San Diego, Big Data Specialization, Coursera, https://www.coursera.org/specializations/big-data.

Co-funded by the Erasmus+ Programme of the European Union

Pytania

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union