Podstawy uczenia maszynowego

Pierwszy projekt od A do Z

O mnie

Rostyslav Apostol

Data Scientist w NorthGravity

https://www.linkedin.com/in/apostolros/

apostol.ros@gmail.com

Co potraficie po warsztacie?

- Rozumienie workflow przy realizacji projektów uczenia maszynowego
- Rozumienie podstawowej niezbędnej teorii
- Umiejętności realizacji projektów ML od A do Z :-)
  - Środowisko
  - Biblioteki
  - Sposób implementacji w Pythonie
  - Metryki ewaluacji modeli

Organizacja dnia

Jak działamy?

Przerwy

Lunch

Agenda

1. Podstawowa teoria
2. Kompleksowy projekt z zastosowaniem klasyfikacji

Podstawy teoretyczne

Co to jest uczenie maszynowe?

"Uczenie maszynowe to nauka polegająca na zmuszaniu komputerów do działania bez wyraźnego zaprogramowania." - Andrew Ng

Nie definiujemy reguł dla komputera. Pozwalamy komputerowi nauczyć się poprzez obserwowanie danych.

Typy uczenia maszynowego

Przykłady użycia ML

1. Przewidywanie ceny nieruchomości
2. Analiza wniosków kredytowych (pozytywna czy negatywna decyzja)
3. Przewidywanie zużycia gazu zimą
4. Przewidywanie kierunku zmiany ceny ropy (wzrośnie czy spadnie?)
5. Segmentacja klientów rynku telefonów komórkowych
6. Przewidywanie prawdopodobieństwa kupna produktu B przez klienta, który kupił produkt A

Etapy typowego projektu ML

1. Zdefiniowanie problemu
2. Zbieranie danych
3. Przetwarzanie i przygotowanie danych
4. Trenowania modeli
5. Wybór i finalizacja najlepszego modelu
6. Używanie modelu

Projekt

Kto z Titanica ma szansę przeżyć?

Zrozumienie problemu - klucz do rozwiązania

Wprowadzenie

15 kwietnia 1912 roku podczas rejsu na trasie Southampton – Cherbourg – Queenstown – Nowy Jork, zderzył się z górą lodową i zatonął.

Spośród 2228 pasażerów i załogi „Titanica” zginęło ponad 1500 osób (68%). Przeżyło katastrofę tylko około 730.

Prawdopodobne przyczyny tak dużych strat w ludziach:

- zbyt mała liczba łodzi ratunkowych,
- przepisy o wymogach oddzielenia pasażerów klasy trzeciej od reszty,
- nadmierna prędkości statku.
- inne.

Chwila refleksji ...

Co mogło decydować o tym, że niektórzy przeżyli, a inni nie?

Zdefiniowanie problemu

Stworzyć **model**, który będzie w stanie przewidzieć, **czy osoba przeżyje** podczas katastrofy czy nie.

Przetestowane zostaną **3 algorytmy** i wybrany zostanie model z **największą dokładnością przewidywania.**

Środowisko i wykorzystywane technologie

Język: Python 3

Środowisko: Google Colab

Biblioteki:

- Numpy
- Pandas
- Matplotlib, Seaborn
- Sklearn

Etap 1 - Pobieranie danych

Github link: https://github.com/RostekA/stacjait-ml/blob/master/titanic.csv

Cel etapu?

Zaciągnięcie danych zewnętrznych do pamięci komputera, żeby móc na danych operować i trenować modele

Wejście: plik csv

Wyjście: dataframe (obiekt w pamięci) i podstawowe dane o datasecie

Podstawowe dane o zbiorze danych

1. Sprawdź 5 górnych wierszy tabeli, żeby mieć pojęcie, co w tej tabeli jest.
2. Ile tabela ma kolumn i wierszy?
3. Czy są wartości zduplikowane w datasecie?

Etap 1 - Podsumowanie

1. Obiekt dataframe (df)
2. Podstawowe informacje na temat zbioru danych

Etap 2 - Analiza eksploracyjna

Cel etapu?

- Zrozumieć dane (ilość, kształt, struktura, relacje)
- Wykryć wartości brakujące (*eng.* missing values)
- Wykryć wartości odbiegające (*eng.* outliers)

Analiza eksploracyjna

Wejście:

Obiekt dataframe na podstawie surowego pliku csv

Wyjście:

Informacje na temat:

- struktury danych (rozkład danych),
- wartości brakujących i odbiegających
- sposobów czyszczenia danych

Pytania techniczne

1. Jakie są typy danych w poszczególnych kolumnach?
2. Jakie są podstawowe metryki statystyczne kolumn?
3. Jakie są rozkłady danych?

Pytania merytoryczne

1. Jaki procent pasażerów przeżył?
2. Czy ma płeć wpływ na wskaźnik przeżycia?
3. Czy samotnie podróżujący miał więcej szans, niż pasażerowie z rodzinami?
4. Czy klasa podróży definiuje szansę na przeżycie?
5. Czy cena biletu w ramach tej samej klasy decyduje o przeżyciu?
6. Czy młodsi mają więcej szans, niż starsi?
7. Czy dzieci mają więcej szans, niż dorośli?
8. Czy miasto, skąd odpływają pasażerowie, przekłada się na wskaźnik przeżycia?

Wartości brakujące i odbiegające

1. Czy są wartości brakujące? Ile?
2. Czy są wartości znacznie odbiegające lub podejrzane?

Wartości brakujące

Algorytmy działają na liczbach, nie na braku liczb.

Metody postępowania z wartościami brakującymi:

- Usuwanie wierszy z przynajmniej jedną wartością brakującą
- Usuwanie kolumn, gdzie mamy zbyt dużo wartości brakujących
- Uzupełnienie wartości brakujących

Jak uzupełniać wartości brakujące?

Zmienna kategoryczna:

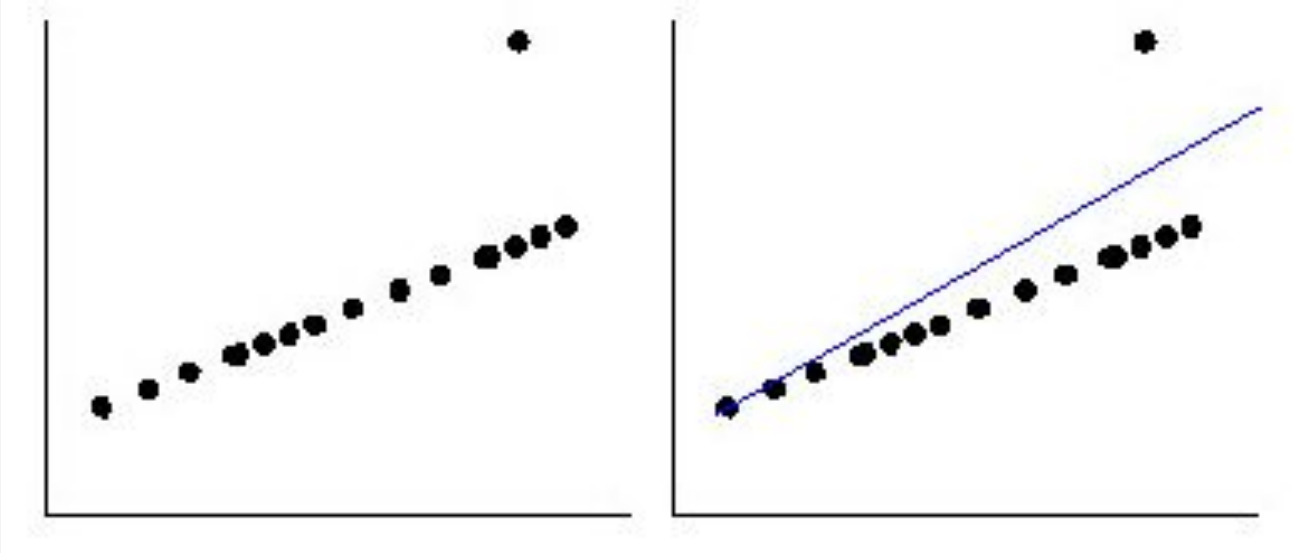
- Stworzyć nową klasę “Missing” / “Brak”
- Uzupełnić wartością modalną
- Uzupełnić na podstawie wiedzy domenowej lub wewnętrznej klasyfikacji

Wartość numeryczna:

- Uzupełnić zerami - wskazuje na brak
- Zamienić wartością średnią lub medianą
- Uzupełnić na podstawie wiedzy domenowej lub wewnętrznej regresji

Wartości odbiegające (Outliers)

Zakłócają modele i obniżają dokładność. Występują w kolumnach numerycznych.



Gdzie jest granica pomiędzy outlier a nie outlier?

IQR (interquartile range) = P(75) - P(25)

Dolna_granica = P(25) - 1,5*IQR

Gorna_granica = P(75) + 1,5*IQR

Przykład:

Miesięczne zarobki brutto programistów = [10, 11, 12, 13, 14, 15, 16, 17, 18, 100]
P(25) = 12 P(75) = 17 IQR = 17 - 12 = 5
Dolna_granica = 12 - 1.5 * 5 = 4,5 Gorna_granica = 17 + 1,5 * 5 = 24,5
Wniosek: 100 to outlier, bo się nie mieści w przedziale [4,5; 24,5]

Jak obsłużyć wartości odbiegające?

1) Nic nie robić. Są outliery naturalne.

   Przykład: populacja Chin na tle populacji innych krajów

2) Usunąć wiersz
3) Zamienić na wartość średnią, medianę
4) Obciąć wartości ekstremalne i zamienić na maxima lokalne

   Przykład:

   1, 2, 3, 4, 5, 100 => 1, 2, 3, 4, 5, 5

Etap 2 - Analiza eksploracyjna: Podsumowanie

Co się udało zrobić?

- Zdobyć wiedzę na temat danych i zależności między nimi

Na przykład, płeć i klasa mocno wpływają na stopień przeżycia.

- Wykryć wartości brakujące i uzupełnić je

Na przykład, wartości dla ‘Embarked’ czy ‘Age’.

- Wykryć wartości odbiegające i obsłużyć je

Na przykład, wartości podejrzane dla ‘Age’ czy cena biletu ‘Fare’.

Etap 3 - Feature Engineering

Cel etapu?

- Zmienne tekstowe / kategoryczne przedstawić w postaci numerycznej
- Tworzenie nowych zmiennych

Wejście: dataframe bez wartości brakujących czy outlierów

Wyjście: dataframe z dodatkowymi zmiennymi oraz enkodowanymi wartościami kategorycznymi

Feature Engineering uważany jest za jeden z najbardziej skutecznych sposobów na zwiększenie dokładności modelu.

Nowe zmienne

Przykład:

Na podstawie wieku - czy osoba jest dzieckiem czy osobą dorosłą?

Na podstawie tytułu i imienia - jaki tytuł posiada osoba?

Wartości kategoryczne

Które kolumny to wartości kategoryczne?

City	Beauty	Population
Warsaw	nice	1,7
Krakow	amazing	0,8
Berlin	so-so	3,7
London	nice	8,9

Jak obsłużyć wartości kategoryczne porządkowe

Label Encoding

Przykład:

Klasa: pierwsza, druga, trzecia => 1, 2, 3

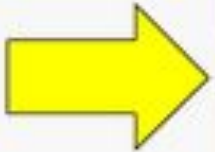
Satysfakcja: źle, przeciętnie, dobrze, super => 0, 1, 2, 3

Jak obsłużyć wartości kategoryczne nie porządkowe

One Hot Encoding

Przykład:

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Etap 3 - Feature Engineering: Podsumowanie

1. Enkodowanie danych tekstowych (Label Encoder czy One Hot Encoder) w zależności od natury zmiennej
2. Tworzenie nowych zmiennych

Etap 4 - Przygotowanie danych

Cel etapu?

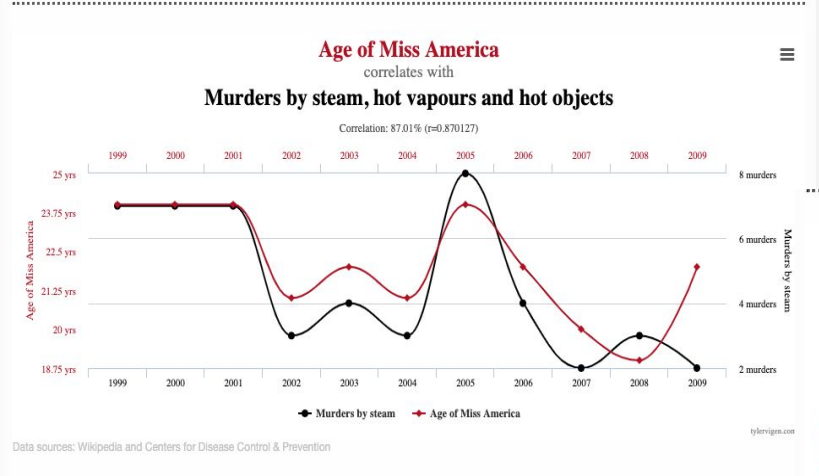
- Feature Selection
- Skalowanie danych
- Podział danych na target i features
- Podział danych na część uczącą i testową

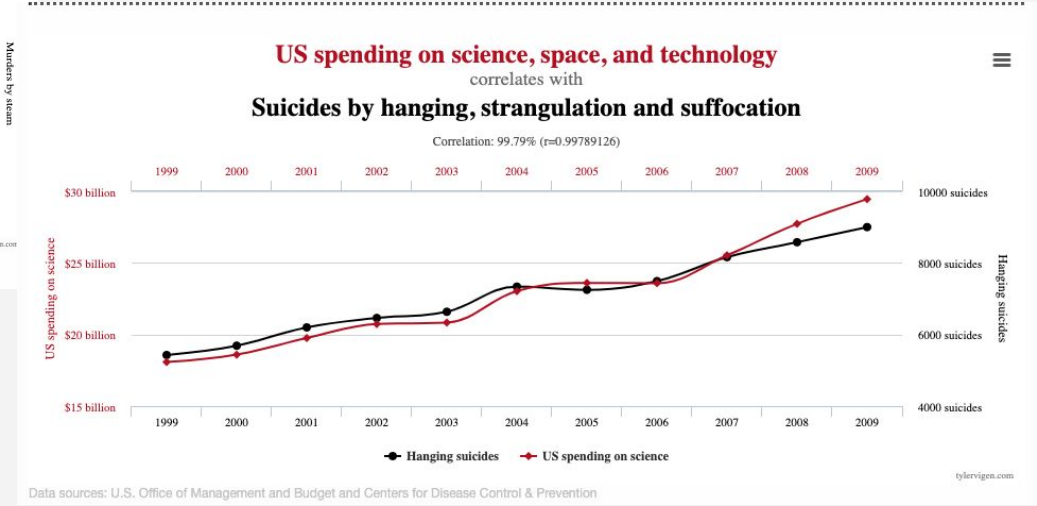
Feature Selection

https://scikit-learn.org/stable/modules/feature_selection.html

1) **Domain based feature selection**
2) Univariate feature selection
3) Recursive feature elimination

Correlation ≠ Causality





Skalowanie

Algorytmy lepiej pracują przy danych w tej samej skali.

Typy skalowania:

- Standardowe skalowanie
- Normalizacja

Optymalnym wyborem jest MinMaxScaler w granicach pomiędzy 0 i 1.

Train test split

Dataset uczący - część zbioru oryginalnego przeznaczona do trenowania modelu.

Dataset testowy - część zbioru oryginalnego, która nie bierze udziału w trenowaniu modelu. Przeznaczona jest do walidacji modelu, czy model działa dobrze na danych, których wcześniej nie widział.

Współczynnik podziału (*split ratio*) - 80/20, 70/30, 90/10.

Etap 4 - Przygotowanie danych: Podsumowanie

- Skalowanie danych z użyciem MinMaxScaler
- Podział danych na target (‘survived’) i features (pozostałe)
- Podział danych na część uczącą i testową ze współczynnikiem 80/20

Etap 5 - Trenowanie modeli

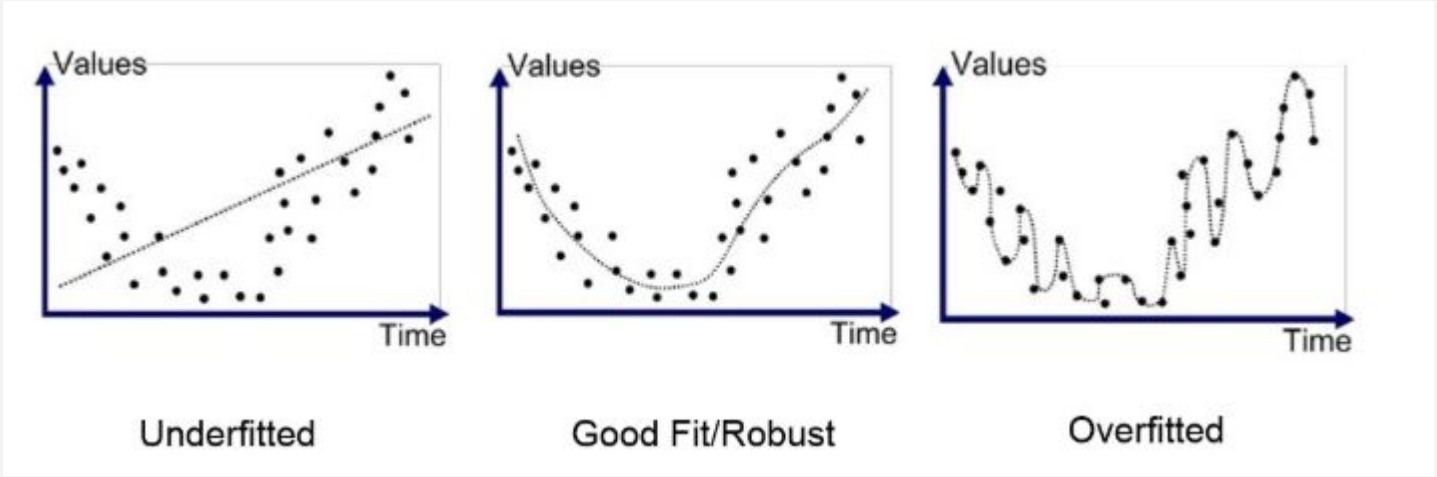
Cel etapu?

Używając algorytmu, próbujemy stworzyć model, który w sposób optymalny odzwierciedla relację pomiędzy zmienną **‘target’** a zmiennymi **‘features’**

Modele:

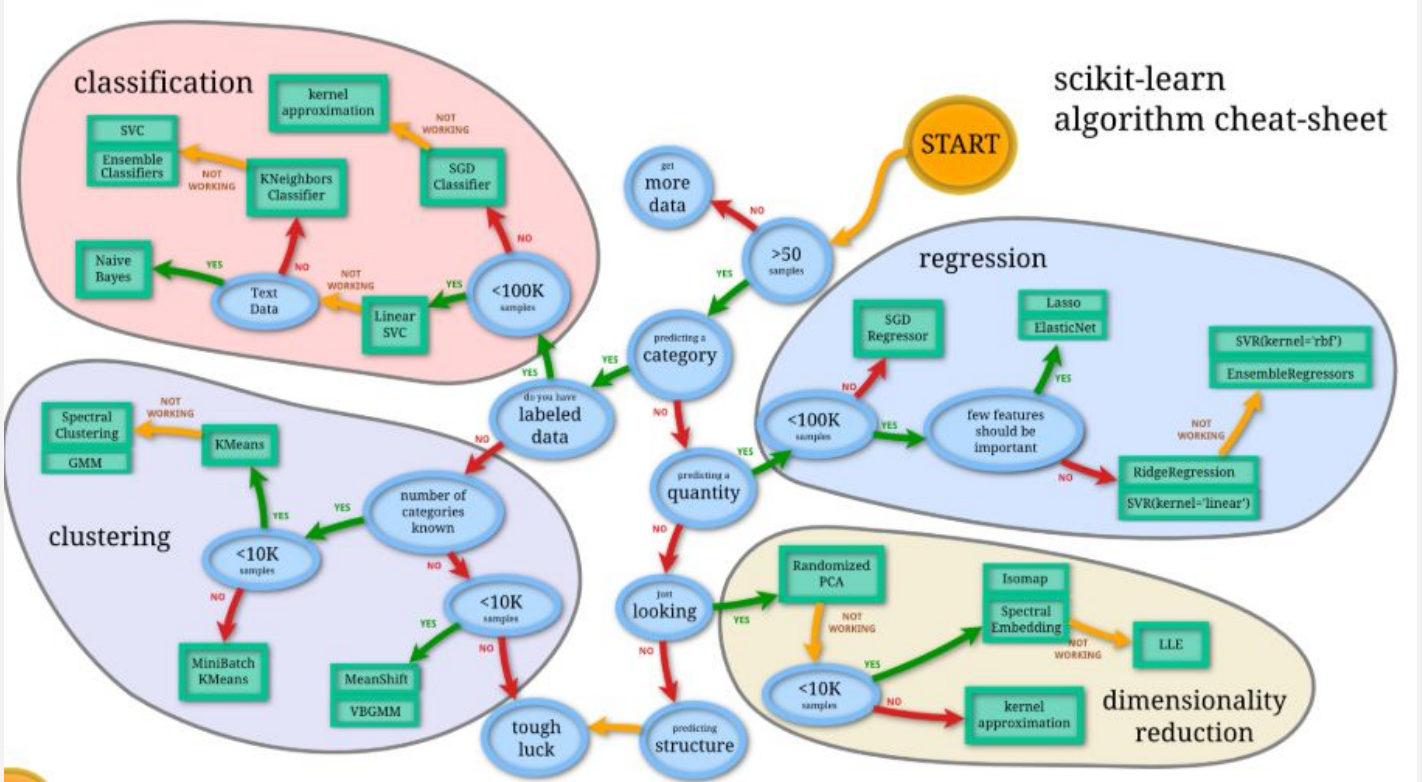
1) Logistic Regression
2) SVM (Supported Vector Machine)
3) K Nearest Neighbors

Overfitting vs. Underfitting

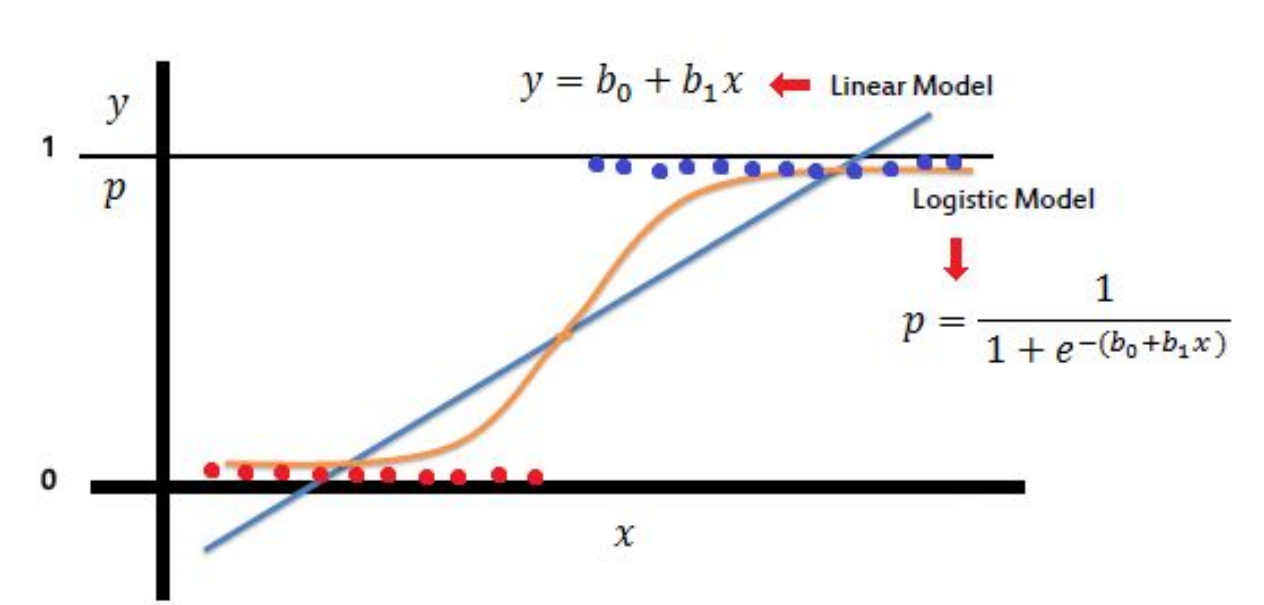


Jak wybrać odpowiedni algorytm?

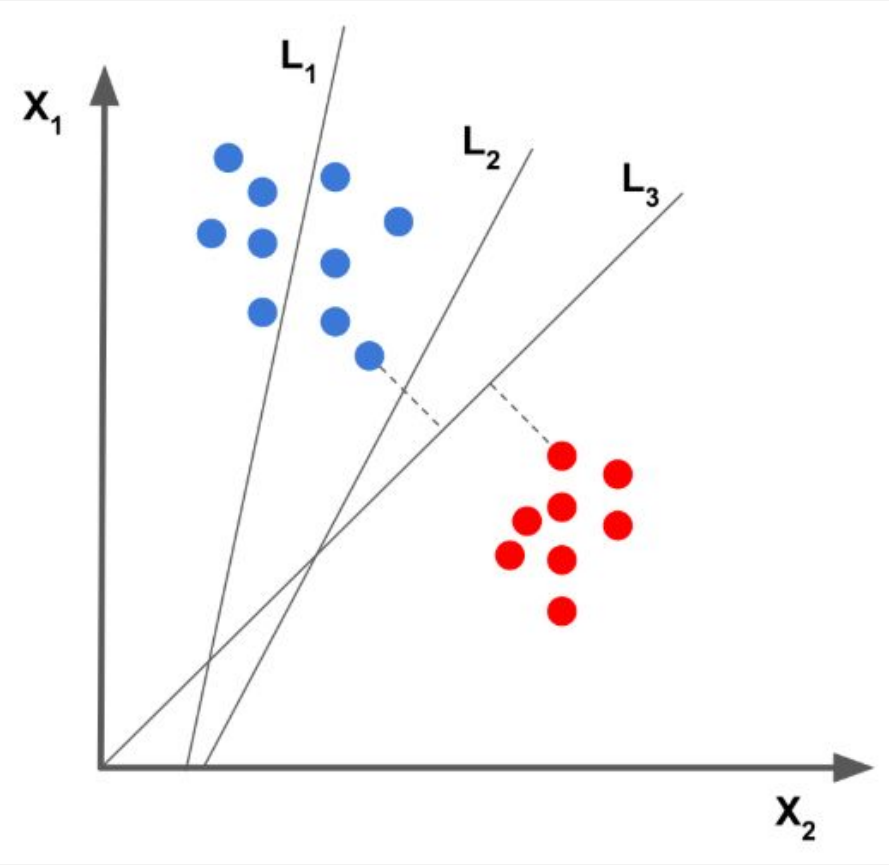
https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html



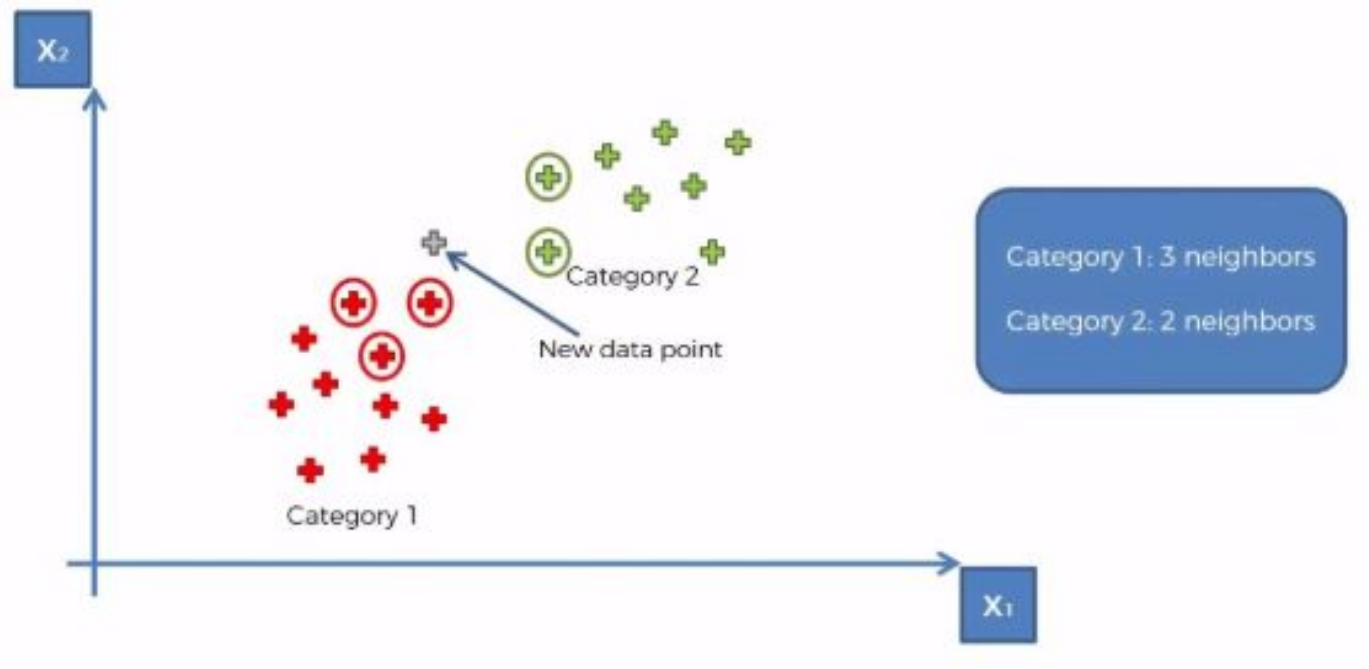
Logistic Regression



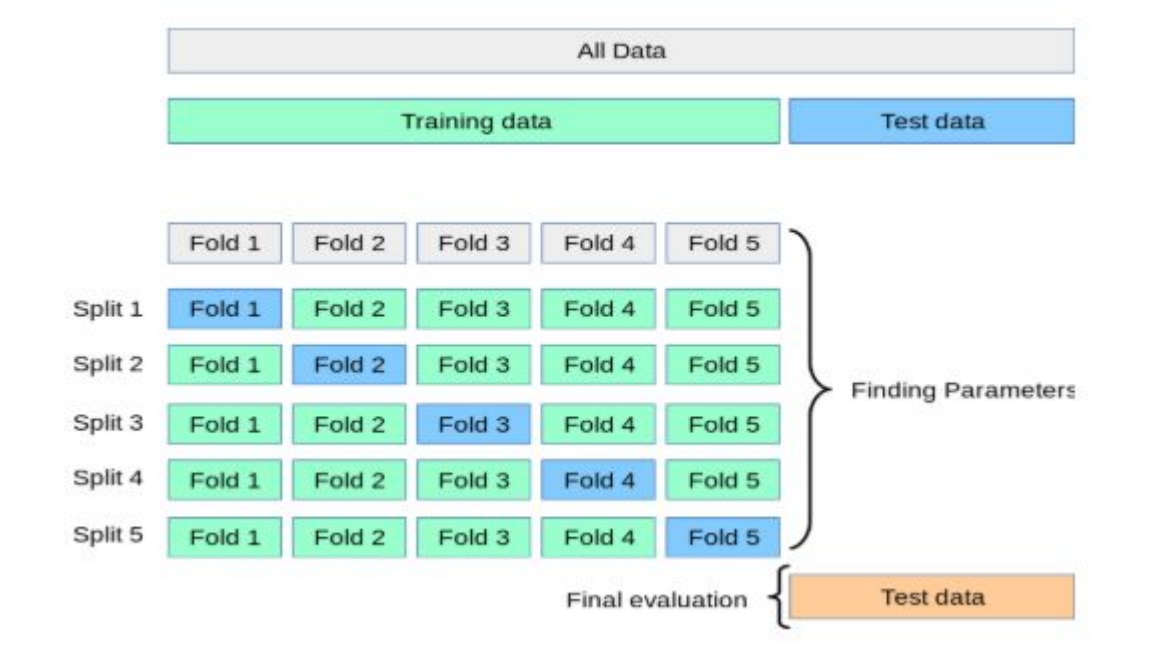
Support Vector Machine



K Nearest Neighbors



Trenowanie i wybór modelu



Metryki ewaluacji modeli

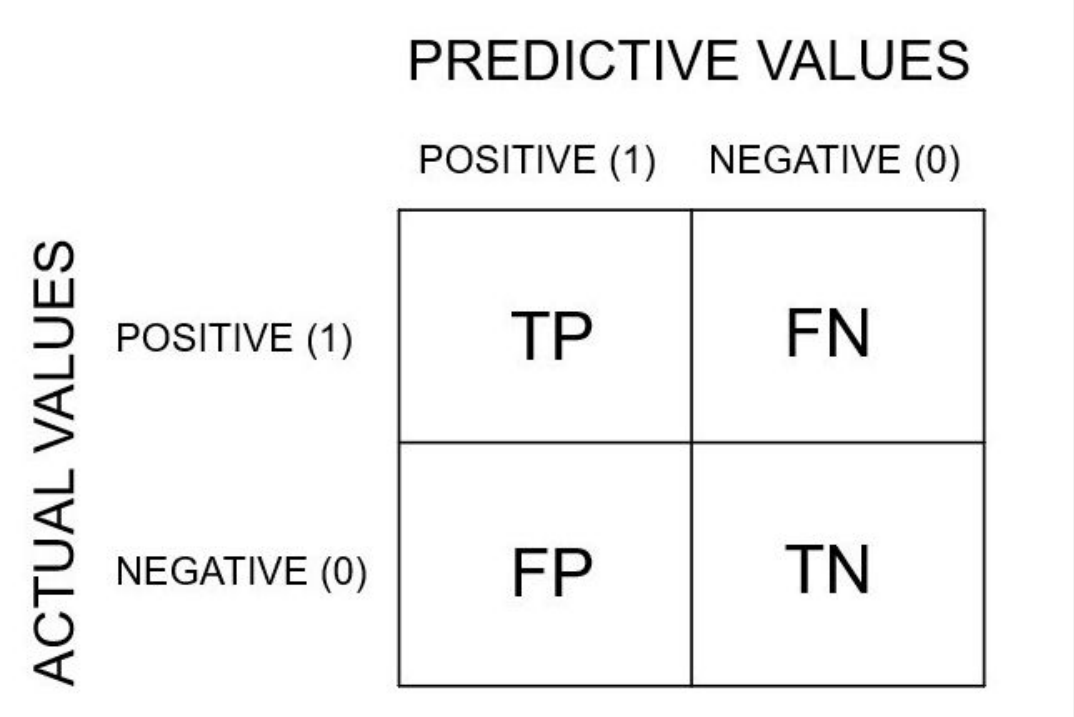
Accuracy = (TP + TN) / (TP +TN + FP + FN)

Precision = TP / (TP + FN)

Recall = TP / (TP + FP)

F1 = 2 * Precision * Recall /

(Precision + Recall)



Wykryć terrorystę - Przykład

Statystycznie jest 10 terrorystów na 100.000 osób.

Etap 5 - Trenowanie modeli: Podsumowanie

- Trzy modele zostały przetrenowane
- Ewaluacja modeli została wykonana

Etap 6 - Finalizacja modelu

Cel etapu?

- Wybór najlepszego modelu
- Trenowanie modelu na całym zbiorze
- Serializacja modelu i wykorzystanie do przewidywania

Podsumowanie projektu

Zrealizowaliśmy takie etapy:

1) Pobieranie danych
2) Analiza eksploracyjna, dane brakujące i outliers
3) Feature Engineering - isChild, isAlone, title
4) Przygotowanie danych - enkodowanie, skalowanie
5) Trenowanie 3 modeli z użyciem kros walidacji
6) Wybór najlepszego modelu - Logistic Regression
7) Przewidywanie z użyciem wytrenowanego modelu

Gdzie szukać więcej informacji?

- Dokumentacja (sklearn, pandas etc.)
- https://machinelearningmastery.com/blog/
- https://towardsdatascience.com
- https://www.kaggle.com
- https://www.datacamp.com/home
- https://www.coursera.org

Zbiory danych

https://towardsdatascience.com/top-10-great-sites-with-free-data-sets-581ac8f6334

Książki teoretyczne

https://www.analyticsvidhya.com/blog/2018/10/read-books-for-beginners-machine-learning-artificial-intelligence/

Pozycje 1,2

https://github.com/rajatmodi62/MLMasteryBooks

Pozycje 3, 6

Dziękuję!

Pytania, sugestie?