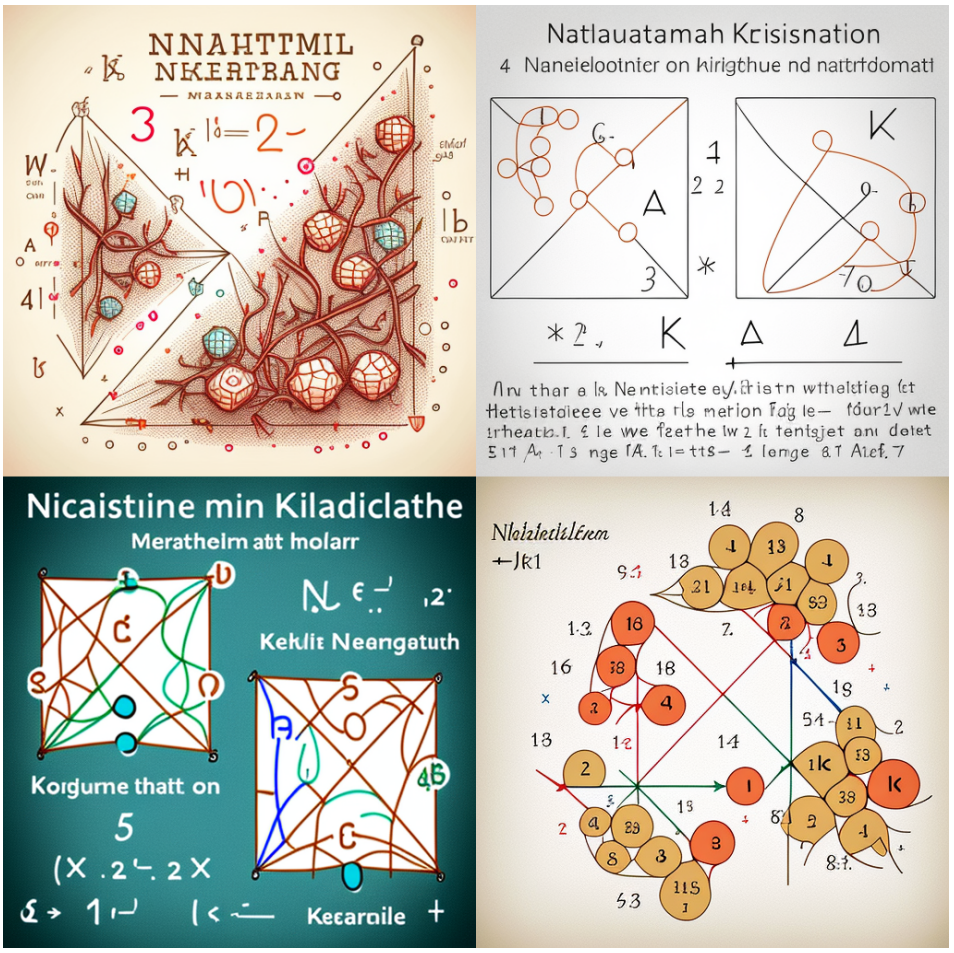
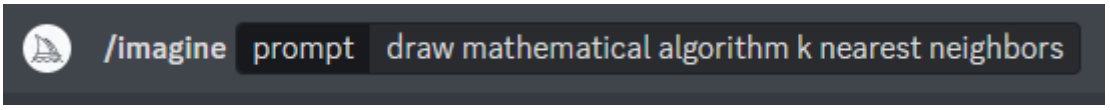
Uważaj, jakie dane pozostawiasz w sieci internetowej - jak sztuczna inteligencja może nas sklasyfikować.



Wygenerowane przez AI - MidJourney wyobrażenie o prezentowanym algorytmie



Autor: Tymoteusz Wszół
Klasa: 7A
Opiekun merytoryczny: nauczyciel matematyki Jarosław Naziemiec
Szkoła Podstawowa nr 2 im. Marii Konopnickiej w Miechowie.
32- 200 Miechów, os. Generała Wł. Sikorskiego 15A tel. sekretariat: (41)383-05-97
https://sp2miechow.szkolnastrona.pl/

Spis treści:

Wstęp **1**

Słownik **2**

Gdzie ta sztuczna inteligencja? **2**

Wprowadzenie do klasyfikacji **4**

Algorytm k-NN **5**

Jak wyznaczyć wartość dla K? 10

Zalety algorytmu k-NN 10

Wady algorytmu k-NN 11

A co w przypadku wartości etykiet, które nie są liczbami rzeczywistymi? 11

Krok po kroku 12

Weryfikacja postawionych tez - jak sklasyfikowałem kolegów **17**

Podsumowanie: i co dalej? **22**

P.S Nie pokazuj tego nauczycielom… **22**

Bibliografia: **23**

Wstęp

Tematem AI zainteresował się w grudniu 2022 roku, kiedy usłyszałem o chat GPT. Zagadnienia AI i ML wydały mi się na tyle interesujące, że postanowiłem przyjrzeć im się bliżej. Zdałem sobie sprawę, że ma to ogromny wpływ na nasze życie, może kształtować nasze zachowania, może być użyte w dobrych jak i złych celach. Nie bez powodu matematyka, nazywana królową nauk, stoi również u podstaw sztucznej inteligencji i nauczania maszynowego. Tematy te są bardzo skomplikowane, ale mam nadzieję, że to do czego udało mi się dojść w matematyce, połączone z moimi przemyśleniami, pozwoli moim kolegom zainteresować się tematem i uświadomić im, że zmiany wykorzystania AI nabrały szybkiego tempa i albo zaczniemy ją rozumieć, albo ona zrozumie nas.

W swojej pracy przedstawię zagadnienie klasyfikacji danych - do czego one służą wraz z przykładami. Skupię się na jednym z nich, obrazując jego praktyczne wykorzystanie, a pomogą mi w nich moi koledzy ze szkoły, dostarczając tym razem świadomie swoich danych do algorytmu, jako danych uczących.

Słownik

Skrót	Ang	Pln
AI	Artificial Intelligence	Sztuczna Inteligencja
ML	Machine Learning	Uczenie Maszynowe
k-NN	. k-nearest neighbor	k najbliższych sąsiadów

Gdzie ta sztuczna inteligencja?

Sztuczna inteligencja to jedno z najciekawszych zagadnień w dzisiejszym świecie. Technologia ta pozwala na stworzenie systemów, które mogą wykonywać zadania wymagające ludzkiej inteligencji czy kreatywności, takie jak rozumienie języka, rozpoznawanie twarzy czy analizowanie danych.

Sztuczna inteligencja jest już wykorzystywana w wielu dziedzinach naszego życia, takich jak medycyna, transport, rolnictwo czy finanse. Przykładowo, w medycynie sztuczna inteligencja pozwala na szybsze i dokładniejsze diagnozowanie chorób, a w transporcie na automatyczne prowadzenie samochodów np: Tesla Elona Muska.

Jednak sztuczna inteligencja niesie również pewne ryzyka. Na przykład, może ona zastąpić ludzi w wielu zawodach, co może prowadzić do ich bezrobocia i niemożności znalezienia pracy zgodnej z ich umiejętnościami. Również jej niekontrolowane rozwijanie może prowadzić do powstawania sztucznych systemów, które przewyższają inteligencją ludzi, co może stanowić dla nich zagrożenie. Nasze wyobrażenia o zagrożeniach z jej strony były już zekranizowane w filmach jak: “Matrix”, “I, Robot”, “Blade Runner” czy nawet “Terminator”.

Mimo tych negatywnych aspektów, sztuczna inteligencja jest jednym z najważniejszych trendów technologicznych współczesności i ma potencjał do poprawy jakości naszego życia. Dlatego ważne jest, aby młodzi ludzie interesowali się tą dziedziną i rozumieli jej możliwości oraz ryzyka, aby przyczyniać się do jej rozwijania w sposób odpowiedzialny i etyczny.

Rola ludzi w świecie, w którym sztuczna inteligencja jest powszechna, będzie się zmieniać w zależności od tego, jakie zadania są przejmowane przez AI i jakie pozostają w rękach ludzi.

AI może zastąpić ludzi w wielu powtarzalnych i rutynowych zadaniach, co może prowadzić do redukcji zatrudnienia w tych obszarach. Jednocześnie jednak, AI może stworzyć nowe możliwości zatrudnienia, np. w dziedzinie projektowania, programowania czy nadzoru. Wydaje się, że będzie to grupa osób, które przynajmniej zrozumieją, jakie podstawy matematyczne za nią stoją.

Ludzie będą również nadal potrzebni do decydowania o tym, jak AI jest wykorzystywana - (ale co z resztą?), ponieważ to ludzie muszą określić cele i zasady działania systemów AI. Będą również odpowiedzialni za rozwiązywanie problemów i kryzysów, które mogą pojawić się w wyniku działania AI.

Ważne jest, aby ludzie nadal rozwijali swoje umiejętności i kompetencje, które nie są łatwe do zastąpienia przez AI, takie jak kreatywność, empatia czy umiejętności interpersonalne. Dzięki temu będą mogli wykorzystać swój potencjał do pracy razem z AI, zamiast zostawać zastąpieni przez nią.

Wspomniałem o pozytywnych aspektach, ale to tę ciemną stronę mocy chcę najbardziej podkreślić. Istnieje wiele przykładów złego wykorzystania sztucznej inteligencji, które mogą mieć negatywny wpływ na różne aspekty naszego życia.

Oto kilka przykładów:

- Manipulacja wyborami: Sztuczna inteligencja może zostać wykorzystana do manipulowania wynikami wyborów poprzez kontrolowanie treści, które są widoczne dla wyborców lub przez fałszowanie danych.

- Biały ekstremizm: AI może być wykorzystana do rozpowszechniania treści rasistowskich czy ksenofobicznych, poprzez automatyczne generowanie treści czy też wykorzystanie algorytmów do promowanie treści ekstremistycznych.

- Dyskryminacja: AI może być zaprogramowana z błędami lub nieetycznie, co może prowadzić do nierównego traktowania ludzi na podstawie ich rasy, płci, orientacji seksualnej czy pochodzenia etnicznego.

- Inwigilacja: AI może być wykorzystywana do inwigilacji obywateli poprzez automatyczne rozpoznawanie twarzy, analizowanie danych z mediów społecznościowych czy też monitorowanie ruchu użytkowników w sieci.

- Bezpieczeństwo: AI może być wykorzystana do ataków hakerskich czy też do przejęcia kontroli nad systemami krytycznymi dla bezpieczeństwa państwa czy infrastruktury.

Dlatego ważne jest, aby odpowiedzialnie rozwijać i wykorzystywać sztuczną inteligencję, z zachowaniem zasad etycznych i zgodnie z prawem. Powinniśmy stale monitorować i regulować jej wykorzystanie, aby zapobiegać negatywnym konsekwencjom i zapewnić bezpieczeństwo i dobro wszystkim.

Matematyka odgrywa kluczową rolę w sztucznej inteligencji. Wiele metod i algorytmów stosowanych w AI, takich jak uczenie maszynowe, sieci neuronowe czy optymalizacja, opierają się na matematycznych modelach i równaniach.

Uczenie maszynowe, jedna z podstawowych dziedzin AI, opiera się na różnych metodach statystycznych i matematycznych, takich jak regresja liniowa, drzewa decyzyjne czy algorytmy genetyczne. Te metody pozwalają na wykorzystanie danych do nauki i przewidywania.

Sieci neuronowe, które są jednym z kluczowych narzędzi w AI, również opierają się na matematycznych modelach. Składają się z wielu warstw, które są połączone ze sobą za pomocą różnych funkcji matematycznych, pozwalających na przetwarzanie danych wejściowych i generowanie odpowiedzi.

Optymalizacja, która jest stosowana w celu znajdowania najlepszych rozwiązań dla różnych problemów, również opiera się na matematycznych metodach, takich jak programowanie dynamiczne, algorytmy genetyczne czy metoda gradientu. Te algorytmy będę poznawał, jak opanuję chociaż jeden język programowania (polecam Pythona).

W związku z tym matematyka jest podstawowym fundamentem AI, gdyż pozwala na modelowanie problemów, tworzenie algorytmów i ich optymalizację, a także na interpretowanie i analizowanie danych.

Każdy nasz ruch w sieci - każde kliknięcie, polubienie jakiegoś materiału zostawia ślad w sieci. Dane te odpowiednio przetworzone przez AI mogą opisać Cię z dokładnością, z jaką znają Cię rodzice i wystarczy ponoć tylko kliknięcie 100 polubień! (jako pierwszy badania na ten temat opublikował Polak Michał Kosiński, patrz nr 4). Być może egzaminy wstępne nie będą już potrzebne w przyszłości, bo AI właśnie na podstawie danych, jakie zostawiliśmy w sieci sama odpowie.

Ale nie martw się, najważniejsze jest to, na co masz wpływ i jesteś tego w pełni świadomy. Teraz koniec tego wyjaśnienia, przyjrzyjmy się bliżej temu, jak to się dzieje.

Wprowadzenie do klasyfikacji

Algorytmy klasyfikacji to metody sztucznej inteligencji, które polegają na przypisywaniu obiektów do określonych klas lub grup. Wiem brzmi strasznie, ale zaraz to wyjaśnię. Służą one do rozpoznawania i zaklasyfikowania danych na podstawie ich cech. Algorytmy klasyfikacji są jednym z najważniejszych elementów uczenia maszynowego i są szeroko stosowane w różnych dziedzinach, takich jak medycyna, bezpieczeństwo, finanse, marketing i wiele innych.

Pisząc prościej: mamy próbkę wielu danych np. długość i szerokość tysięcy gruszek i jabłek. Klasą lub grupą są właśnie jabłka i gruszki. Jak na podstawie znajomości szerokości i długości 1001 obiektu odpowiedzieć na pytanie czy jest to jabłko i gruszka?

Tutaj pomoże nam statystyka.

Teraz wyobraź sobie, że mamy setki odpowiedzi uczniów na kilka tematów, np: jakie lubią filmy, w jakie gry lubią grać, czy interesują się sportem. I teraz na podstawie tych danych należy określić prawdopodobieństwo, czy inni uczniowie przekroczą wymagany próg punktów z egzaminu z matematyki?

Algorytm k-NN

Na początku trochę teorii. Może wydawać się nużąca ale pokaże jej działanie na dość zabawnym przykładzie.
Kroki algorytmu prezentują się następująco:

1. Zdefiniuj zbiór treningowy $D = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$,
   gdzie x_n to wektor (*vector*) cech (np. odpowiedzi na jakieś pytania), a " y_n to etykieta (*label*) (klasa, np. odpowiedź na szukane pytanie).

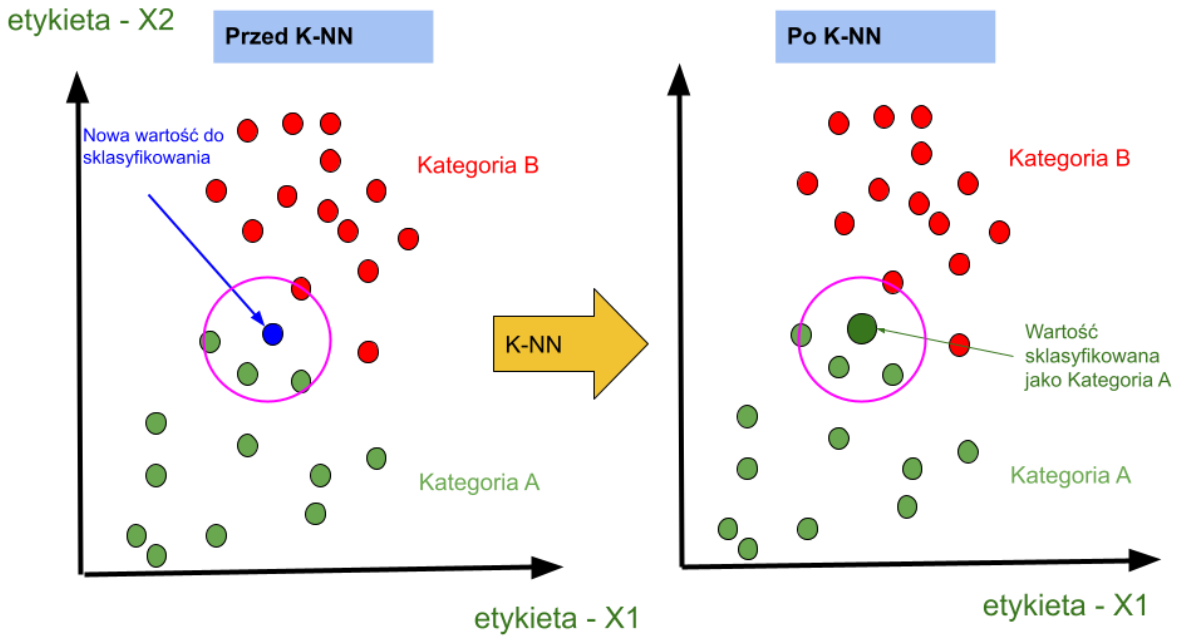
2. Dla każdego nowego przykładu x, który chcemy sklasyfikować, oblicz odległość euklidesową (o tym później) między x a każdym przykładem z D.

3. Wybierz k najbliższych przykładów z D do x.

4. Przypisz x do klasy, która jest najczęściej reprezentowana w, k-a najbliższych przykładach.

5. Wartość k jest hiperparametrem, który jest ustawiany przed trenowaniem modelu. Co oznacza, że jest wartością stałą ustaloną przed uruchomieniem algorytmu, nazwa dla stałej ale tak nazwana w obszarze AI.

Jako, że rysunek powie więcej niż słowa, zobaczmy na przykładzie wizualizacji jak dokonuje się taka klasyfikacja:



Rys. 1 Zobrazowanie działania algorytmu k-NN w przestrzeni 2 wymiarowej

Etykieta X1 obrazuje odpowiedź na zadane pytanie nr 1, etykieta X2 na pytanie nr 2. Ale jak to? Wyżej w przykładzie była odpowiedź tak/nie!

Tak, ale teraz zastanów się jak przedstawić to jako wartość liczbową w zbiorze liczb rzeczywistych?

Możemy przypisać odpowiedzi tak wartość 1 a nie = 0. Algorytm potrzebuje liczbowej reprezentacji.

Wobec czego dla skończonego zbioru odpowiedzi np. A, B, C, D, E itd. możemy przypisać konkretną wartość liczbową. Są oczywiście wyjątkowe przypadki, do czego wrócimy później, na razie możesz zastanowić się, jakie to mogą być przypadki.

Kolor obrazuje nam wynik klasyfikacji, czyli przyporządkowanie do danej kategorii. Np. kategoria A - to uczeń wyróżniający się a kategoria B - to uczeń wymagający większej uwagi. Punkt niebieski obrazuje nam pojawienie się nowych wartości dla X1, X2, dla których poszukujemy kategorii - w naszym przypadku odpowiedzi do jakiej klasy przypisać nowego ucznia.

K - tak jak opisywałem to z góry określona liczba przez nas. Przyjmij, że jest to liczba 4. Wobec czego szukamy najbliższych 4 sąsiadów naszego punktu. Najbliższych mam na

myśli takich do, których jest najmniejsza odległość. 3 na 4 należy do kategorii A, wobec czego prawdopodobieństwo, że nowy uczeń będzie należał do kategorii A jest większe i wynosi ¾ = 75%. Wiem, że to było proste, ale teraz zapiszmy w sposób ogólny coś, co nazwiemy wzorcem, przepisem albo właśnie krokami algorytmu:

Krok-1 Przypisz wartość zmiennej K - liczba najbliższych sąsiadów.

Krok-2 Oblicz odległość Euklidesową (o coś nowego) od nowego punktu od pozostałych.

Krok-3 Wybierz K sąsiadów, którzy są najbliżej nowego punktu.

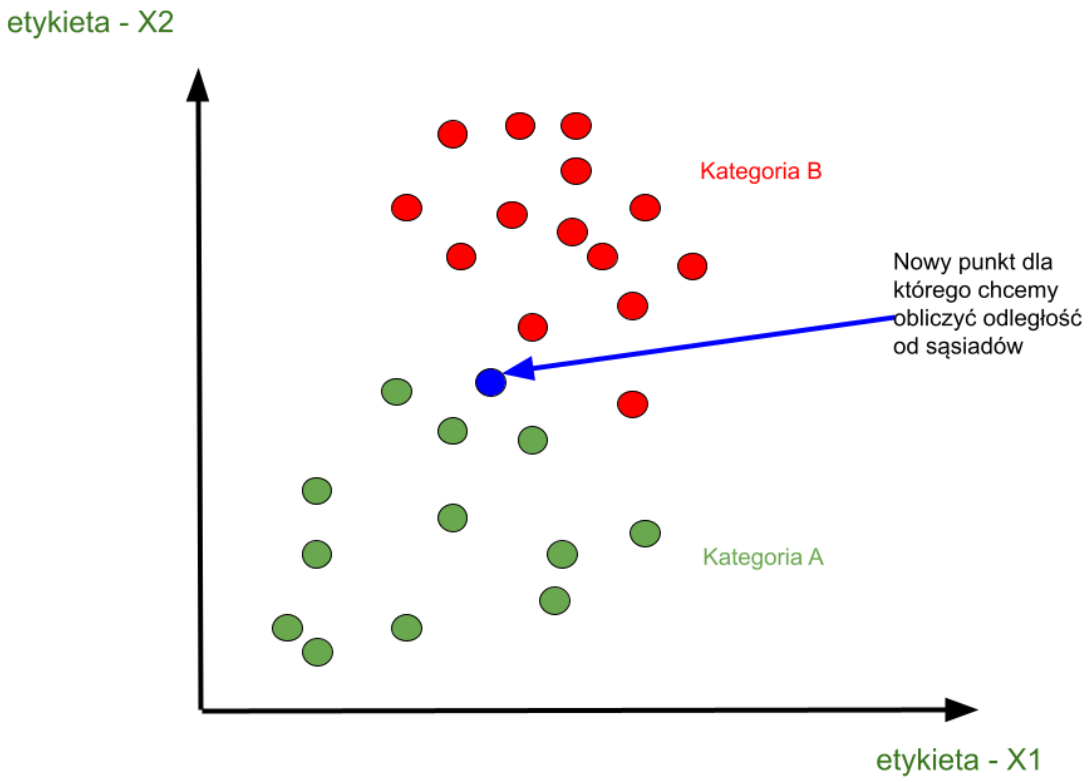
Krok-4 Spośród punktów wyznaczonych w kroku nr 3 policz ile należy do poszczególnej kategorii.

Krok-5 Przypisz nowy punkt do kategorii, która ma najwięcej punktów wyznaczonych w kroku nr 4.

Krok-6 Koniec, mamy już sklasyfikowany nowy punkt.

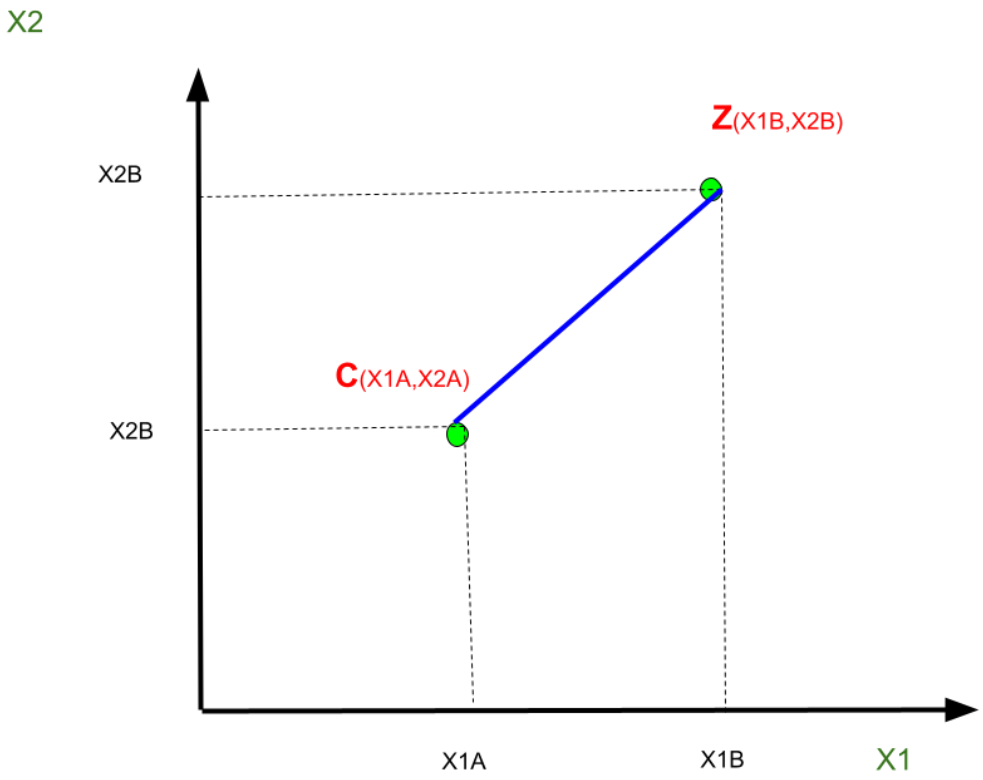
Proste prawda?

A teraz czym jest ta odległość Euklidesowa i co zrobić, jeśli etykiet będzie więcej niż 2 czy 3.



Rys. 2 Zobrazowanie działania algorytmu k-NN w przestrzeni 2 wymiarowej

Na rysunku mamy punkt zaznaczony kolorem niebieskim. Teraz chcemy wyznaczyć odległości od pozostałych punktów.



Rys. 3 Odległość punktów w przestrzeni 2-wymiarowej

Odległość Euklidesowa między punktem C i Z w przestrzeni 2-wymiarowej:

$$d(C,Z) = \sqrt{(X1_B - X1_A)^2 + (X2_B - X2_A)^2}$$

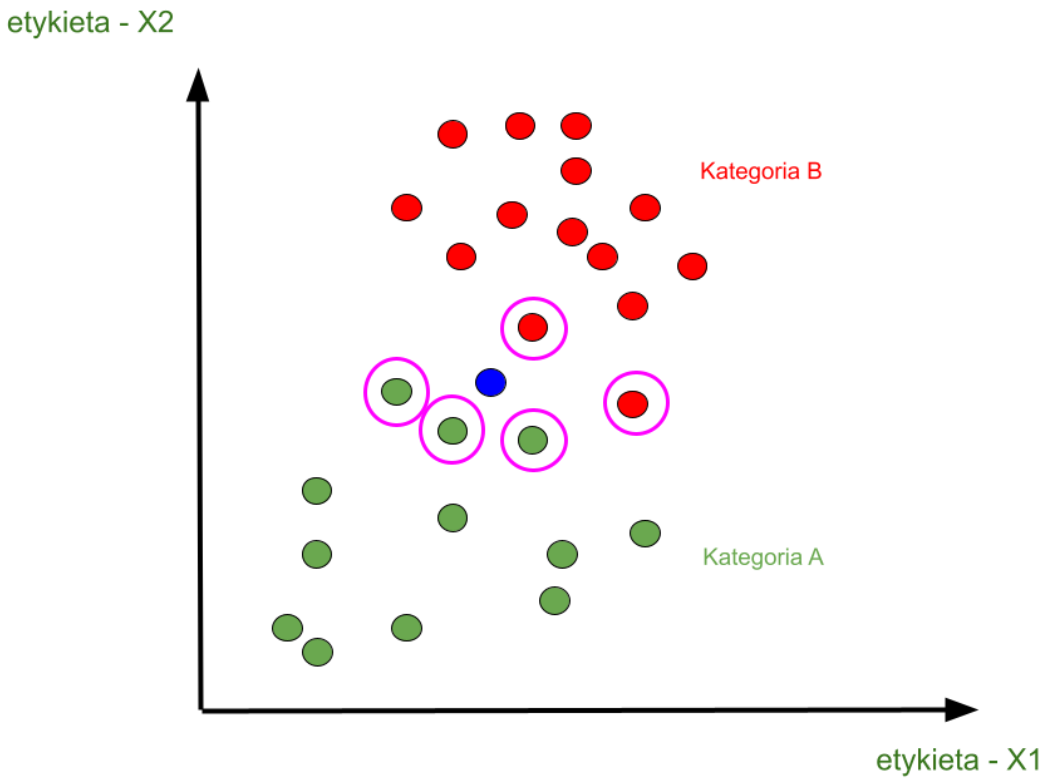
w przestrzeni 3-wymiarowej:

$$\sqrt{(X1_B - X1_A)^2 + (X2_B - X2_A)^2 + (X3_B - X3_A)^2}$$

i w przestrzeni n-wymiarowej.

$$d(C,Z) = \sqrt{\sum_{n=1}^{\infty} (Xi_B - Xi_A)^2}$$

Jesteś w stanie wyobrazić sobie przestrzeń 4-wymiarową? Ja też nie byłem, ale zachęcam do poszukiwania w internecie.



Rys. 4 5 najbliższych sąsiadów

Dla K=5 czyli pięciu najbliższych sąsiadów wyznaczamy liczbę punktów należących do danej kategorii. I tak 3 należą do kategorii i A do kategorii B.

Nowy punkt zostaje zaklasyfikowany do kategorii A, gdyż 3 z 5 najbliższych sąsiadów stanowi większość.

Jak wyznaczyć wartość dla K?

- Algorytm nie określa jaka wartość K będzie prawidłowa. Możesz wypróbować kilka różnych wartości K i porównać wyniki. W literaturze zaleca się zacząć od K=5.

- Wartości jak K=1 i K=2 mogą doprowadzić do mylących wniosków.

Zalety algorytmu k-NN

- Jest prosty w zrozumieniu
- Daje miarodajne wyniki dla danych losowych
- Dla dużych danych treningowych czas obliczeń jest przewidywalny

Wady algorytmu k-NN

- Trzeba z góry znać wartość K
- Koszt obliczeń jest stosunkowo duży, bo wymaga przeliczenia odległości między elementami zbioru danych, dlatego też algorytmy takie nazywane są lazy (leniwymi), bo potrzebują całego zbioru danych wejściowych, aby móc przeprowadzić klasyfikację.

A co w przypadku wartości etykiet, które nie są liczbami rzeczywistymi?

Dobre pytanie. Postaram się na nie odpowiedzieć.

W takim przypadku musimy każdej z wartości - jaka by ona nie była - przypisać wartość liczbową.

I tak dla skończonego zbioru odpowiedzi zestaw liczb:

- dla tak/nie - wartość 0,1 albo 1,3
- dla tak/nie/nie wiem - wartość 0, 0.5 , 1 albo 1,2,3
- dla skończonego zbioru odpowiedzi jak kolor
  biały, niebieski, zielony … czarny - 1,2,3 .. 30

Odpowiedzi w postaci liczb nie musimy przekształcać, np. jaki masz wzrost i ile ważysz.

Teraz zapewne zastanawiasz się co ma piernik do wiatraka. Jak kolor można porównać ze wzrostem. Odpowiedź jest prosta: tak można. A co z jednostkami?

Jak km/h mogę porównać z km albo z kg?

A co jak jedno pytanie będzie zawierało odpowiedzi tak i nie, do których przypiszemy wartość 0 i 1, a drugie pytanie będzie o wzrost w nanometrach?

W takim przypadku najlepiej dane te przeskalować do tego samego przedziału np. 0-1,

tak, by zniwelować zbyt duże rozbieżności między nimi.

I jak zwykle mamy na to przepis.

Aby zaprezentować wszystkie dane na tej samej skali od [0,1] możemy użyć do tego normalizacji min-max.

$$x^{'} = \frac{x - min(x)}{max(x) - min(x)}$$

jak zwykle najlepiej posłużyć się przykładem. Załóżmy, że mamy zbiór odpowiedzi jak [1,2,40,60,200] dane po przeniesieniu na skalę od 0 do 1 będą miały wartość:

$1^{'} = \frac{1-1}{200-1} = 0$ a więc wartość 1 zaprezentujemy jako 0,

$40^{'} = \frac{40-1}{200-1} \simeq 0.196$ wartość 40 w przybliżeniu jako 0.196

$60^{'} = \frac{60-1}{200-1} \simeq 0.296$ wartość 60 w przybliżeniu jako 0.296

$200^{'} = \frac{200-1}{200-1} = 1$ wartość 200 w przybliżeniu jako 1

Teraz już masz komplet wiedzy. Wiesz jak zamienić każdą odpowiedź na wartość liczbową i jak dane te przeskalować i zaprezentować na wspólnej skali.

Krok po kroku

Przed eksperymentem, policzymy bardziej skomplikowany przykład.
Załóżmy, że grupie 10 uczniów zadajemy 4 pytania. 3 z nich posłużą nam jako wektor a 4 będzie przypisaniem do konkretnej klasy.

Przykładowe pytania i odpowiedzi to: (odpowiedzi wygenerowałem samodzielnie):

1. **Czy uprawiasz sport?**
2. **Czy lubisz słodycze?**
3. **Czy jadasz warzywa?**
4. **Twój stan zdrowia**

Możliwe odpowiedzi na pytania 1-3 to elementy z listy:

tak
nie
czasami
rzadko
w ogóle nie

Możliwe odpowiedzi do określenia stanu zdrowia to:

bardzo dobry
dobry
dostateczny
mierny

Odpowiedzi prezentują się następująco:

	Czy uprawiasz sport?	Czy lubisz słodycze?	Czy jadasz warzywa?		Twój stan zdrowia
Uczeń 1	tak	nie	czasami		bardzo dobry
Uczeń 2	tak	czasami	tak		bardzo dobry
Uczeń 3	tak	rzadko	czasami		dobry
Uczeń 4	czasami	czasami	czasami		dobry
Uczeń 5	rzadko	rzadko	w ogóle nie		dostateczny
Uczeń 6	nie	tak	nie		mierny
Uczeń 7	nie	czasami	w ogóle nie		mierny
Uczeń 8	rzadko	nie	tak		dostateczny
Uczeń 9	nie	rzadko	czasami		dostateczny
Uczeń 10	rzadko	rzadko	czasami		?

Uczeń 10 odpowiada tylko na 3 zadane pytania. Naszym zadaniem jest określenie jego stanu zdrowia, czyli przypisanie do klasy z pytania 4-tego.

Kolejnym krokiem jest przypisanie wartości do elementów ze zbioru odpowiedzi i ich znormalizowanie, tak by była możliwość porównania różnych zbiorów między sobą i określenie odległości.

i tak:

Odpowiedzi	Przyjęta wartość liczbowa (może być inna, ale ja przyjąłem taką)	Normalizacja min-max
tak	100	$\frac{100-0}{100-0} = 1$
nie	10	$\frac{10-0}{100-0} = 0.1$
czasami	50	$\frac{50-0}{100-0} = 0.5$
rzadko	30	$\frac{30-0}{100-0} = 0.3$
w ogóle nie	0	$\frac{0-0}{100-0} = 0$

Odpowiedzi	Przyjęta wartość liczbowa (może być inna, ale ja przyjąłem taką)	Normalizacja min-max
bardzo dobry	100	$\frac{100-0}{100-0} = 1$

dobry	80	$\frac{80-0}{100-0} = 0.8$
dostateczny	50	$\frac{50-0}{100-0} = 0.5$
mierny	0	$\frac{0-0}{100-0} = 0$

Teraz nasze dane możemy przedstawić w postaci liczbowej tak:

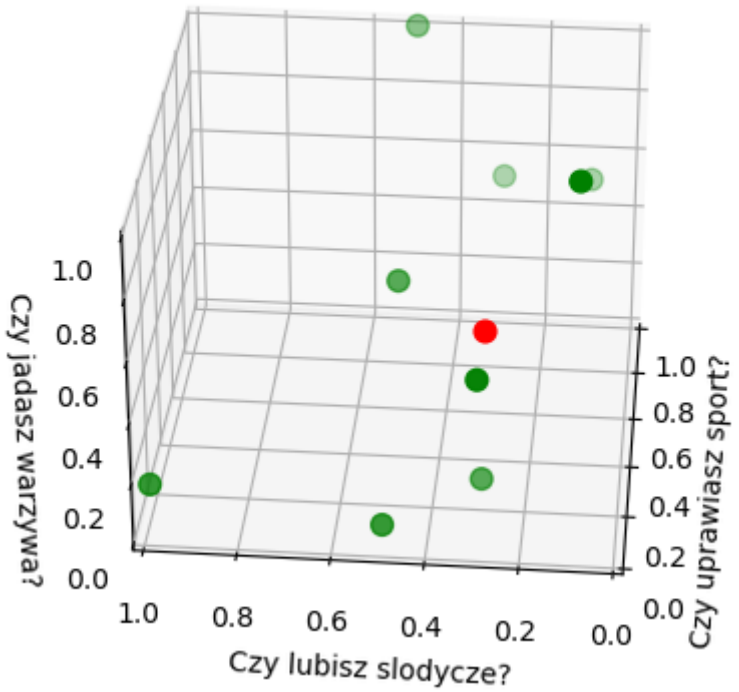
	Czy uprawiasz sport?	Czy lubisz słodycze?	Czy jadasz warzywa?		Twój stan zdrowia
Uczeń 1	1,00	0,10	0,50		1,00
Uczeń 2	1,00	0,50	1,00		1,00
Uczeń 3	1,00	0,30	0,50		0,80
Uczeń 4	0,50	0,50	0,50		0,80
Uczeń 5	0,30	0,30	0,00		0,50
Uczeń 6	0,10	1,00	0,10		0,00
Uczeń 7	0,10	0,50	0,00		0,00
Uczeń 8	0,30	0,10	1,00		0,50
Uczeń 9	0,10	0,30	0,50		0,50
Uczeń 10	0,30	0,30	0,50		?

Kolejnym krokiem jest obliczenie odległości Euklidesowej wektorów Ucznia 10 od pozostałych:

	Odległość Ucznia 10 od:
Uczeń 1	$\sqrt{(X1_B - X1_A)^2 + (X2_B - X2_A)^2 + (X3_B - X3_A)^2} =$ $\sqrt{(1 - 0.3)^2 + (0.1 - 0.3)^2 + (0.5 - 0.5)^2} =$ $\sqrt{0.7^2 + 0.2^2 + 0^2} =$ $\sqrt{0.7^2 + 0.2^2 + 0^2} =$ $\sqrt{0.49 + 0.04 + 0} =$ $\sqrt{0.53} \simeq 0.73$
Uczeń 2	0.88
Uczeń 3	0.70
Uczeń 4	0.28
Uczeń 5	0.50
Uczeń 6	0.83
Uczeń 7	0.57
Uczeń 8	0.54
Uczeń 9	0.20

Uczeń 10	0,00

Zweryfikuj czy Twoje wyniki pokrywają się z moim, dla ułatwienie rozpisałem krok po kroku jak obliczyć odległości między Uczeń 10 a Uczeń 1.



Rys. 5 Reprezentacja wektorów w przestrzeni 3 wymiarowej.

Żeby zwizualizować odległośći narysowałem dla Ciebie ich reprezentację w przestrzeni 3 wymiarowej. Czerwony punkt to nasz Uczeń 10. (Python i biblioteka do rysowania matplotlib)

Jak pamiętasz K dla algorytmu określamy z góry, ale sprawdźmy jak będą różnić się wyniki dla pierwszych 5 wartości.

Dla K=1.

Najbliższy sąsiad jest w odległości 0.2 od Uczeń 10. Jest to uczeń 9. Jako, że uczeń 9 należy do klasy dostateczny. To nasz Uczeń 10 zostaje przypisany do klasy również dostateczny z prawdopodobieństwem 100%.

K=1	Praw. K=1
bardzo dobry	0,00%
dobry	0,00%
dostateczny	100,00%
mierny	0,00%

Dla K=2.

Najbliżsi dwaj sąsiedzi są w odległości 0.2 (Uczeń 9 dostateczny) i 0.28 (Uczeń 4 dobry). 1 uczeń z 2 reprezentuje klasę dostateczną i 1 uczeń 2 reprezentuję klasę dobrą.

Wobec czego:

K=2	Praw. K=2
bardzo dobry	0,00%
dobry	50,00%
dostateczny	50,00%
mierny	0,00%

Odpowiednio dla K 3,4,5 dostajemy:

K=3	Praw. K=3	K=4	Praw. K=4	K=5	Praw. K=5
bardzo dobry	0,00%	bardzo dobry	0,00%	bardzo dobry	0,00%
dobry	33,33%	dobry	25,00%	dobry	20,00%
dostateczny	66,67%	dostateczny	75,00%	dostateczny	60,00%
mierny	0,00%	mierny	0,00%	mierny	20,00%

Też otrzymałeś takie wyniki? Jeśli tak, to gratuluję, dostajesz dodatkowy punktu w drodze do zrozumienia AI.

Nasunęły Ci się jakieś wnioski? Mi tak więc się nimi podzielę:

1. Musimy posiadać dużo wiarygodnych danych, w przypadku gdyby wszyscy byli zdrowi, bez względu na odpowiedzi z poprzednich pytań, każda nowa odpowiedź również byłaby zakwalifikowana jako zdrowa.

2. Wyniki mogą się różnić dość znacząco w zależności od wybranego K.

3. Algorytm dla każdego nowego pytania musi liczyć odległości względem elementów z całego zbioru, im większy zbiór, tym dłużej będzie to trwało.

Weryfikacja postawionych tez - jak sklasyfikowałem kolegów

A teraz przejdźmy do eksperymentu. Te same pytania zadałem kolegom ze szkoły.

Więc dane wejściowe są stosunkowo bardziej “realne”.

Wyliczenia przedstawione w pkt “Krok po kroku” zautomatyzowałem przenosząc do excela, dzięki temu wynik dostaję automatycznie (link tutaj: excel).

Oto wyniki:

	Czy uprawiasz sport?	Czy lubisz słodycze?	Czy jadasz warzywa?		Twój stan zdrowia	Dystans do Ucznia i
Uczeń 1	tak	czasami	rzadko		bardzo dobry	0,71
Uczeń 2	nie	rzadko	rzadko		dobry	0,81
Uczeń 3	tak	rzadko	nie		dobry	0,88
Uczeń 4	nie	nie	nie		dobry	1,00
Uczeń 5	tak	tak	tak		bardzo dobry	0,86
Uczeń 6	w ogóle nie	rzadko	tak		dobry	1,11
Uczeń 7	w ogóle nie	tak	nie		dostateczny	0,54
Uczeń 8	rzadko	rzadko	rzadko		bardzo dobry	0,73
Uczeń 9	tak	czasami	rzadko		dobry	0,71
Uczeń 10	czasami	nie	tak		bardzo dobry	1,14
Uczeń 11	czasami	tak	rzadko		dobry	0,00
Uczeń 12	tak	tak	rzadko		dobry	0,50
Uczeń 13	czasami	nie	rzadko		dobry	0,90
Uczeń 14	nie	rzadko	rzadko		dobry	0,81
Uczeń 15	tak	czasami	tak		bardzo dobry	0,99
Uczeń 16	tak	czasami	rzadko		dobry	0,71
Uczeń 17	tak	czasami	nie		bardzo dobry	0,73
Uczeń 18	nie	rzadko	tak		dobry	1,07
Uczeń 19	rzadko	w ogole nie	nie		dobry	1,04
Uczeń 20	czasami	czasami	w ogole nie		dobry	0,58
Uczeń 21	tak	w ogole nie	w ogole nie		dobry	1,16
Uczeń 22	czasami	tak	czasami		bardzo dobry	0,20
Uczeń X	czasami	tak	rzadko			0,00

Tabela 1. Odpowiedzi uczniów. Uczeń X to szukana klasyfikacja

	Czy uprawiasz sport?	Czy lubisz slodycze?	Czy jadasz warzywa?		Twój stan zdrowia	Dystans do Ucznia i
Uczeń 1	1,00	0,50	0,30		1,00	0,71
Uczeń 2	0,10	0,30	0,30		0,80	0,81
Uczeń 3	1,00	0,30	0,10		0,80	0,88
Uczeń 4	0,10	0,10	0,10		0,80	1,00
Uczeń 5	1,00	1,00	1,00		1,00	0,86
Uczeń 6	0,00	0,30	1,00		0,80	1,11
Uczeń 7	0,00	1,00	0,10		0,50	0,54
Uczeń 8	0,30	0,30	0,30		1,00	0,73
Uczeń 9	1,00	0,50	0,30		0,80	0,71
Uczeń 10	0,50	0,10	1,00		1,00	1,14
Uczeń 11	0,50	1,00	0,30		0,80	0,00
Uczeń 12	1,00	1,00	0,30		0,80	0,50
Uczeń 13	0,50	0,10	0,30		0,80	0,90
Uczeń 14	0,10	0,30	0,30		0,80	0,81
Uczeń 15	1,00	0,50	1,00		1,00	0,99
Uczeń 16	1,00	0,50	0,30		0,80	0,71
Uczeń 17	1,00	0,50	0,10		1,00	0,73
Uczeń 18	0,10	0,30	1,00		0,80	1,07
Uczeń 19	0,30	0,00	0,10		0,80	1,04
Uczeń 20	0,50	0,50	0,00		0,80	0,58
Uczeń 21	1,00	0,00	0,00		0,80	1,16
Uczeń 22	0,50	1,00	0,50		1,00	0,20
Uczeń X	0,50	1,00	0,30		#N/A	0,00

Tabela 2. Odpowiedzi uczniów zamienione na wartości liczbowe.

Odpowiedzi	**Wartość liczbowa**	**Normalizacja min-max**		**Odpowiedz**	**Wartość liczbowa**	**Normalizacja min-max**	
tak	100	1,00	tak	bardzo dobry	100	1,00	bardzo dobry
nie	10	0,10	nie	dobry	80	0,80	dobry
czasami	50	0,50	czasami	dostateczny	50	0,50	dostateczny
rzadko	30	0,30	rzadko	mierny	0	0,00	mierny
w ogóle nie	0	0,00	w ogóle nie				

Tabele 3. Normalizacja zbioru odpowiedzi do wartości od 0-1.

K=1	Praw. K=1	K=2	Praw. K=2	K=3	Praw. K=3	K=4	Praw. K=4	K=5	Praw. K=5
bardzo dobry	0,00%	bardzo dobry	50,00%	bardzo dobry	33,33%	bardzo dobry	25,00%	bardzo dobry	20,00%
dobry	100,00%	dobry	50,00%	dobry	66,67%	dobry	50,00%	dobry	60,00%
dostateczny	0,00%	dostateczny	0,00%	dostateczny	0,00%	dostateczny	25,00%	dostateczny	20,00%
mierny	0,00%	mierny	0,00%	mierny	0,00%	mierny	0,00%	mierny	0,00%

Tabela 4. Wyniki na różnych wartości K.

Podsumowanie: i co dalej?

Na tym nie koniec, mam nadzieję, że zaciekawiłem Was wstępem do AI. Od czegoś trzeba zacząć - zagadnienie na pierwszy rzut oka wydaje się skomplikowane, ale jak dobrze poznamy aparat matematyczny, jaki za nim stoi, będziemy lepiej go rozumieć.

Warto w kolejnych krokach poznać kolejne algorytmy klasyfikacji. W 8 klasie pojawia się temat związany z prawdopodobieństwem, a tu już mały kroczek do naiwnej klasyfikacji Bayesa. Pewnie już znasz Paradoks Monty'ego Halla? Jeśli nie to, zachęcam do zapoznania, świetnie obrazuje, jak intuicja może nas zawieść. Bayes nie tylko pozwala nam uciec z pętli paradoksu, ale również można wykorzystać go to klasyfikacji.

To nie koniec, więcej możliwości poznawczych zyskasz, gdy poznasz lepiej nie tylko matematykę, ale również programowanie. Ja już wyruszyłem w tę drogę, to dopiero początek. Jedno co wiem, to że muszę zwracać baczniejszą uwagę na to, jaki ślad po sobie zostawiam w internecie, bo jak opisałem - nieznanych zagrożeń w tej przygodzie może zostać odkrytych wiele.

P.S Nie pokazuj tego nauczycielom…

AI, która zastępuje proces twórczy malarstwa - https://midjourney.com/

AI, które pomoże Ci w procesie nauki https://chat.openai.com/

Bibliografia:

1. Sztuczna inteligencja, Janiszewski Boguś, Skorwider Max.
2. https://pl.wikipedia.org/wiki/K_najbli%C5%BCszych_s%C4%85siad%C3%B3w
3. https://chat.openai.com/
4. https://midjourney.com/
5. https://www.pnas.org/doi/10.1073/pnas.1218772110
6. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
7. https://pl.frwiki.wiki/wiki/Hyperparam%C3%A8tre
8. https://pl.wikipedia.org/wiki/Przestrze%C5%84_euklidesowa
9. https://pl.wikipedia.org/wiki/Paradoks_Monty%E2%80%99ego_Halla
10. https://en.wikipedia.org/wiki/Feature_scaling#Rescaling_(min-max_normalization)

Miechów, 20.02.2023r.

Opinia

Zakres materiału realizowanego w szkole podstawowej z matematyki i informatyki jest dalece niewystarczający do zajęcia się tak złożonymi problemami jak poruszone przez Tymoteusza w jego pracy. Dlatego doceniam jego determinację i poświęcenie niezbędne do zgłębienia wiedzy stojącej u podstaw tworzenia sztucznej inteligencji. Tymoteusz w swojej pracy korzystał z ogólnie dostępnych w Internecie publikacji, które wykorzystał tworząc własne przykłady mogące spopularyzować poruszony problem wśród rówieśników.

W związku z powyższym pozytywnie opiniuję pracę ucznia i oczekuję, że będzie ona dla Niego początkiem wspaniałej przygody z naukami ścisłymi.

nauczyciel matematyki

mgr inż. Jarosław Nazimek