Wykrywanie występowanie chorób serca,porównanie algorytów uczenia maszynowego nadzorowanego na podstawie zbioru danych dotyczących chorób układu krążenia z repozytorium UCI

Magdalena Szulc

2022-05-01

Contents

Abstrakt

The aim of the work is to compare selected algorithms of supervised machine learning and build a model based on medical data, which diagnoses the presence or absence of cardiovascular disorders.

Medical data is distinguished by the fact that it is difficult to access it, most often it is not information that is made available for public use, therefore, a key step is to choose the features taken into account when creating the model. The data obtained from the UCI repository has already undergone preliminary processing, the dataset itself, due to its small size, allows checking effects of algorithms without getting rid of redundant and insignificant features.

The main motive is to answer the question of how data deficiency strongly influences the outcome and whether there is a difference between the use of selected supervised learning algorithms requires a comparison of the ease of creating a model, accuracy, complexity and time to obtain an answer.

Wykrywanie występowanie chorób serca,porównanie algorytów uczenia maszynowego nadzorowanego na podstawie zbioru danych dotyczących chorób układu krążenia z repozytorium UCI

Wstęp

Sztuczna inteligencja wśród szerokiego zakresu swoich zastosowań może zostać wykorzytana do analizy bardziej lub mniej złożonych danych medycznych, w celu przewidzenia wystąpienia choroby u konkretnej osoby, bez udziału procesu myślowego od stony specjalisty.

Do tego przeznaczenia istnieje możliwość zastosowania uczenia nadzorowanego (ang. *supervised learning*) tj. rodzaj uczenia maszynowego zakładający istnienie zbioru danych testowych zawierających odpowiedzi, na których podstawie wyszukiwane są zależności, cechy znaczące oraz budowany jest w ten sposób model służący przykładowo do przewidywania przyszłych wartości.

W przypadku danych dotyczących chorób zależności typujące występowanie choroby, bazują na podstawie konkretnych wyników badań zgormadzonych w repozytorium UCI.

W dzisiejszych czasach choroby sercowo-naczyniowe stanowią najczęstszą przyczynę zgonów, a liczba osób cierpiących na te dolegliwości stale rośnie. Głównymi przyczynami zachorowalności diagnozowanymi przez specjalistów są niski poziom świadomości i profilaktyki chorób serca.Dlatego prowadzone są intensywne prace nad zwiększeniem dostępności badań, które wspomogą diagnostykę kardiologiczną na jak najwcześniejszym etapie [1].

Powodem szukania dokładniejszych sposobów diagnozowania są również wysokie koszty leczenia generowane przez choroby układu krwionośnego. Według analityków firmy konsultingowej KPMG [2] w 2011 r. koszty diagnostyki i terapii chorób serca wyniosły ponad 15 miliardów polskich złotych.

Uczenie maszynowe poprzez przetwarzanie dużych zasobów klinicznych danych historycznych pod kątem zależności przyczynowo skutkowych, może zostać wykorzystane do wczesnej diagnostyki lub wspomagania leczenia pacjentów [3].

Słowa kluczowe: uczenie maszynowe, uczenie nadzorowane, lasy losowe, maszyna wektorów nośnych, k-najbliższych sąsiadów

[1] Wojciech Modrzejewski and Włodzimierz J. Musiał tyt.:"Stare i nowe i czynniki ryzyka sercowo-naczyniowego - jak zahamować epidemię miażdżycy? Część I. Klasyczne czynniki ryzyka", Forum Zaburzeń Metabolicznych 2010;1(2):106-114.

[2] międzynarodowa sieć firm audytorsko-doradczych ze szczególnym uwzględnieniem branży dóbr konsumpcyjnych, usług finansowych, nieruchomości i budownictwa, technologii informacyjnych, mediów i komunikacji (TMT), transportowej (TSL), produkcji przemysłowej, a także sektora publicznego

[3] Korczak, Karol. "Uczenie maszynowe w opiece zdrowotnej." Roczniki Kolegium Analiz Ekonomicznych/Szkoła Główna Handlowa 56 Technologie informatyczne w administracji publicznej i służbie zdrowia (2019): 305-316.

Cel i zakres pracy

Celem pracy jest porównanie wybranych algorytmów uczenia maszynowego nadzorowanego, przy założeniu że dane wejściowe są wybrakowane, a w rezultacie zbudowanie modelu który na podstawie danych medycznych wystawia diagnozę o występowaniu zaburzeń sercowo-naczyniowych lub ich braku.

Dane medyczne wyróżniają się tym, że trudno uzyskać do nich dostęp, najczęściej nie są to informacje, które się udostępnia do użytku publicznego, z tego powodu, kluczowym krokiem jest wybór cech branych pod uwagę przy tworzeniu modelu. Dane pozyskane z repozytorium UCI przeszły już wstepną obróbkę , sam dataset ze względu na swoje niewielkie rozmiary pozwala na sprawdzenie działań algorytmów bez pozbywania się nadmiarowych i mało znaczących cech.

Zatem odpowiedź na pytanie jak wybrakowanie danych mocno wpływa na rezultat i czy istnieją róznicę między zastosowaniem wybranych algorytmów nauczania nadzorowanego wymaga przedstawienia porówniania łatwości tworzenia modelu, dokładności, złożoności oraz czasu uzyskania odpowiedzi.

W pracy opisano następujące algorytmu uczenia nadzorowanego:

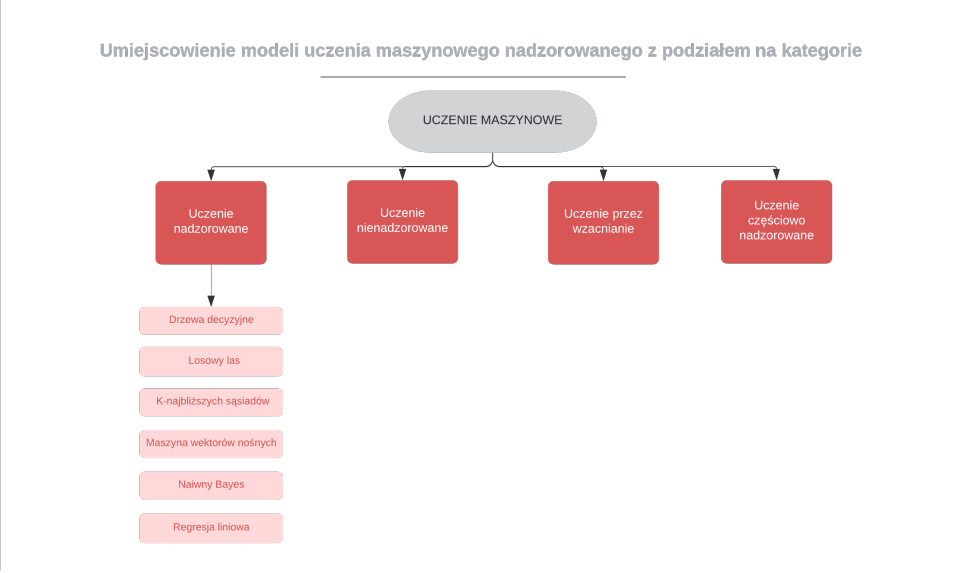
- losowe lasy decyzyjne (ang. *random decision forests*)
- maszyna wektorów nośnych (ang. *support vector machines*, SVM)
- k-najbliższych sąsiadów (ang. *k-neares neighbours*, KNN)

Wprowadzenie teorertyczne

Uczenie maszynowe (ang. *machine learning*, ML) to dziedzina zajmująca się tworzeniem modeli do analizy bardzo obszernych zasobów danch. Modele utworzone za pomocą algorytmów uczenia maszynowego są w stanie z wysokim prawdopodobieństwem wystawić predyckję lub dokonać klasyfikacji na temat zadanego problemu.

Model *klasyfikacjny* służy do przewidzenia etykiety klasy poprzez mapowanie na już z góry ustalony jednowymiarowy pozdział, model *regresywny* natomiast mapuje przestrzeń ustalając liczbę klas podziału oraz grupując wartości. [4] Istnieje możliwość przekształcenia problemu regresywnego na klasyfikacyje i na odwrót poprzez zamiane wartości oczekiwanego wyniku. Taką modyfikację zastosowano w praktycznej częsci projektu. Wyniki dla danych występowały w wartościach od 0 do 4 , dla wartości <1,4> przypadek testowy uznawany był za sklasyfikowany pozytywny (chory), dlatego przekształcenie z modelu regresywnego do modelu klasyfikacyjnego polega na konwersji wyników do wartości liczbowych 0 - brak stwierdzenia stanu choroboweo oraz 1 - stwierdzenie o chorobie układu krążenia.

Sposób wykorzystania segreguje alorytmy uczenia maszynowego na dwie kategorie, jednak powszechnie stosowanym podziałem jest podział zależnie od sposobu *trenowania* algorytmu. Algorytmy dzieli się na min.: uczenie nadzorowane, uczenie częściowo nadzorowane, uczenie bez nadzoru oraz uczenie przez wzmacnianie [5] .



Dobór typu uczenia oraz algorytmu uzależniony jest od danych wejściowych oraz oczekiwanego rezultatu. Dane wyjściowe mogą przyjmować format odpowiedzi TAK/NIE , klasyfikacji do danego zbioru czy np procentowej oceny ryzyka.

Uczenie maszynowe nadzorowane (ang. *supervised learning*) to klasa algorytmów uczenia maszynowego, która bazuje na poetykietowanych danych. Nadzór polega na porównaniu rezultatów działania modelu z wynikami które są zawarte w danych wejściowych (*dane oznaczone*) [6]. Algorytm po osiągnięciu żądanej efektywnosci jest w stanie dokonać klasyfikacji

[4] An overview of the supervised machine learning methods Vladimir Nasteski Faculty of Information and Communication Technologies, Partizanska bb, 7000 Bitola, Macedonia

[5] Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

[6] The use of machine learning methods in predicting stock prices on the stock exchange. Autor: Tomasz Łysiak

przykładu dla którego nie posiada odpowiedzi. Sprawdza się to obecnie w rekomendacji produktów oraz diagnozie chorób. Z matematycznego puntu widzenia dopasowanie danych oznaczonych nazywane jest aproksymacją funkcji [7] .

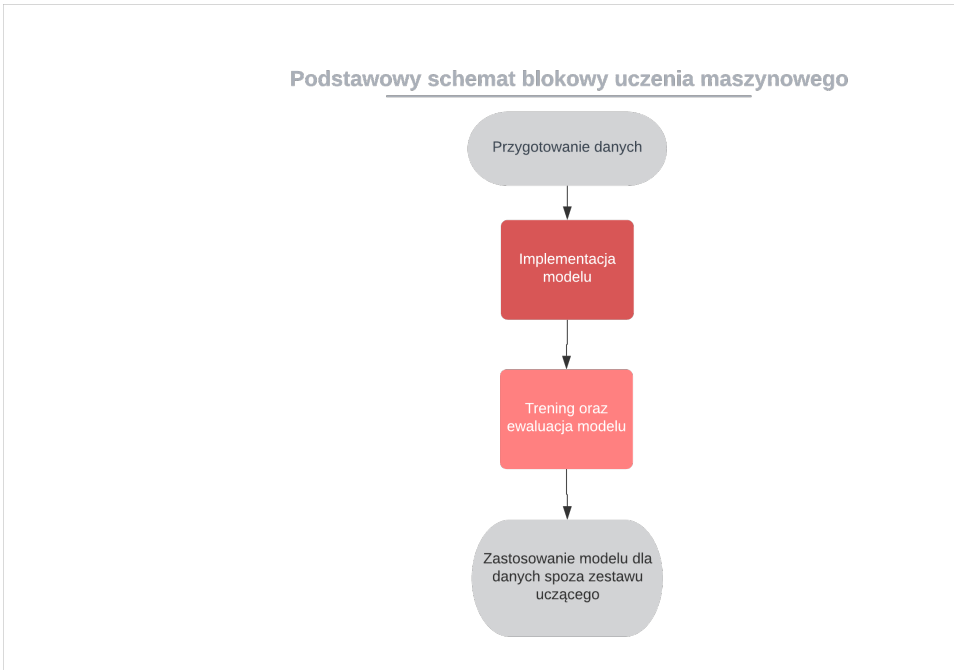
Uczenie maszynowe bez nadzoru (ang. *unsupervised learning*) to klasa algroytmów uczenia maszynowego która wiodąco rozwiązuje problemy grupowania. Dane dostarczane do modelu nie zawierają *oznaczeń*, zatem nauczanie polega na wyciąganiu konkluzji z poprzednio wykonanych iteracji. Na skuteczność modeli budownych w oparciu o uczenie bez nadzoru wpływ ma rozmiar dostarczonego do nauki zbioru danch, im jest on większy tym bardziej wzrasta effektywność. Takie zbiory można uzyskać rejestrując dane na bierząco dlatego do najczęstrzych zastosowań tej klasy algorytmów, można zaliczyć rozpoznawanie mowy czy obrazu [8] .

Uczenie maszynowe przez wzmacnianie (ang. *reinforcement learning*) to klasa algorytmów uczenia maszynowego której nauczanie nie opiera się na danych wejściowych czy wyjściowych a rezultatach otrzymanych podczas testu nazywanych tzw. sygnałami wzmocnienia który może przyjmować wartość pozytwną lub negatywną. Algorytm generując dane wejściowe dostosowuje reguły by uzyskać zwrotnie sygnał pozytywny w jak największej liczbie przypadków. [9] .

Uczenie częściowo nadzorowane (ang. *semi-supervised learning*) to klasa algorytmów uczenia maszynowego która wykorzystuje zbór danych w większości niepoetykietowany na podstawie których tworzony jest model [10] .

Podział osób na kategorie cierpiące na choroby sercowo-naczyniwe oraz zdrowe, to dylemat klasyfikayjny nadający się do rozwiązania za pomoca algorytmów uczenia maszynowego nadzorowanego i na nich skupia się dalsza część pracy.

Ścieżka działania algorytmów uczenia maszynowego nadzorowanego



[7]Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

[8]Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

[9]An Overview of Machine Learning Methods Used in Sentiment Analysis. Justyna Laska

[10]van Engelen, J.E., Hoos, H.H. A survey on semi-supervised learning. Mach Learn 109, 373–440 (2020). https://doi.org/10.1007/s10994-019-05855-6

Model Danych

[11]

Repozytorium uczenia maszynowego UCI

Sensem wykorzystania uczenia maszynowego jest prognoza lub klasyfikacja rzeczywistych wartości z dużego zbioru danych które mogą znaleść zastosowanie w praktycznych dziedzinach. Im bardziej dokładne i rzeczywiste dane do testowania i tworzenia modelu tym większe prawdopodobieństwo otrzymania realnych wyników na końcu ścieżki uczenia. W celu gromadzenia miarodajnej bazy dostępnych zbiorów danych testowych powstało repozytorium uczenia maszynowego UCI. Jak podaje strona informacyjna :

> … było ono cytowane ponad 1000 razy, co czyni je jednym ze 100 najczęściej cytowanych „artykułów" w całej informatyce … [12]

Repozytorium gromadzi dane z wielu rozbieżnych dziedzin , dane medyczne umieszczone w repozytorium nie zawierają wrażliwych danych pacjentów , a niektóre zbiory są poddane już wstępnej obróbce tak jak zbiór danych "Heart Disease Databases" wykorzystany w tym dokumencie, który powstał na podstawie realnych danych medycznych zebrany z lokalizacji

1. Fundacja Cleveland Clinic [13]
2. Węgierski Instytut Kardiologii, Budapeszt [14]
3. V.A. Centrum medyczne, Long Beach, Kalifornia [15]
4. Szpital Uniwersytecki, Zurych, Szwajcaria [16].

Stratyfikacja

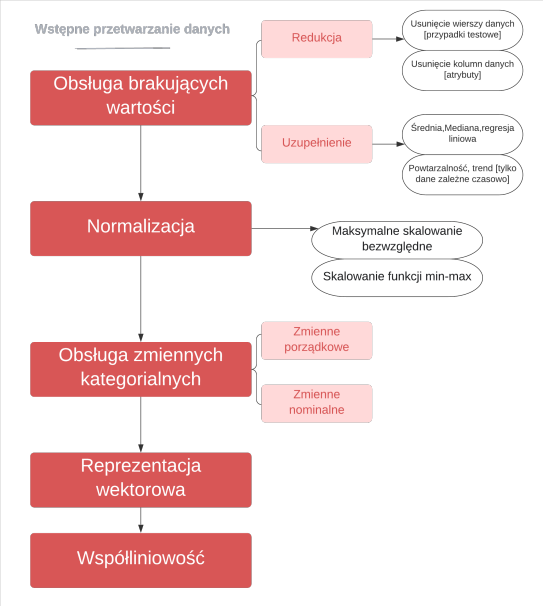
Wyróżniono 14 atrybutów spośród 76 zebrancyh do wykorzystania w algorytmach uczenia maszynowego, wszystkie z nich mają wartości liczbowe.

[Todo dodać jak dzielą się dane na kobiety mężczyżni]

W przypadku danych testowych z repozytorium UCI, fakt iż dane pochodziły z różnych lokalizacji ma duże znaczenie ,gdyż od placówki medycznej zależy jakim badaniom poddani zostali pacjęci a co za tym idzie w jakich kolumnach tabelarycznego przedstawienia będą mieć uzupełnione bądź puste wartości. Scalenie ze sobą wyników badań dostarcza większej różnorodności również dzięki temu że dane pochodzą z wielu krajów. Jeżeli zestaw wejściowy zostałby ograniczony do jednej lokalizacji to cecha dla której nie uzupełniono wartości zostałaby pominięta podczas treningu ze względu na brak danych, co skutowało by uboższym modelem i możliwe że pominięciem kluczowej cechy wpływającej na działanie.

Proces przetwarzania danych może składać sie z wielu różnych kroków zależenie od typu, w uczeniu nadzorowanym operującym na danych tekstowo-liczbowych poprawnym będzie zastosowanie schematu przedstawionego poniżej:

[11] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA:University of California, School of Information and Computer Science.
[12] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA:University of California, School of Information and Computer Science.
[13] V.A. Fundacja Centrum Medyczne, Long Beach i Cleveland Clinic:dr n. med. Robert Detrano
[14] Węgierski Instytut Kardiologii. Budapeszt: Andras Janosi, MD
[15] V.A. Fundacja Centrum Medyczne, Long Beach i Cleveland Clinic:dr n. med. Robert Detrano
[16] Szpital Uniwersytecki, Zurych, Szwajcaria: William Steinbrunn, MD i Szpital Uniwersytecki, Bazylea,Szwajcaria: Matthias Pfisterer, MD



Po złączeniu można przeprowadzić szereg działań w celu sztucznego uzupełnienia pustych wartości bazując na wartościach które już istnieją.

Obsługa brakujących wartości

Możliwościami obsługi brakujących wartości są : mniej polecana ze względu na utrate danych, redukcja zestawu danych lub uzupełnienie go zgodnie z wybrany przez siebie założeniem. Biblioteki do nauczania maszynowego dostarczają już gotowe rozwiązania do upuszczenie wierszy lub kolumn zawierających wartości *null*. Uzupełnienie danych inaczej *imputacja*, rozwiązuje problem w mniej stratny sposób i tak samo jak do redukcji są już gotowe rozwiązania w bibliotece sklearn. Istnieją 4 różne strategie uzupełniania wykorzystujące proste matematyczne obliczenia takie jak :

- średnia,
- mediana,
- stała,
- najczęściej występująca wartość.

Do wyznaczenia wartości uzupełniających można również użyć regresji liniowej.

Standaryzacja

Przeksztaucenie danych również bazujące na statystycznych założeniach i również ustandaryzowane w popularnych bibliotekach. Dążymy aby średnia wartoś wynosiła 0, a odchylenie standardowe 1 dla liczbowyh reprezentacji danych. Z matematyczne punktu widzenia wykonujemy działanie

$$\frac{\bar{X}}{\sqrt{\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{N-1}}}$$

[17]

Obsługa zmiennych kategorialnych

Cechy kategorialne dzielą się na dwie zasadnicze grupy ze względu na mozliwość uprorządkowania , dane takie jak wykształcenie , rozmiar podlegają mapowaniu , dane typu kolor lub płeć podlegaja kodowaniu. W ten sposób dane kategoryczne

[17] Peshawa J. Muhammad Ali, Rezhna H. Faraj; "Data Normalization and Standardization: A Technical Report", Machine Learning Technical Reports, 2014, 1(1), pp 1-6.

staja się wartosciami liczbowymi.

Reporezentacja wektorowa

Obsługa danych kategorialnych pozwoliła zmapować/zakodować je w postaci liczbowej, ale można pójść o krok dalej i te same dane mieć w postaci 0 lub 1 na odpowiedniej kolumnie. Rozwiązanie reprezentacji wektorowej polega na utworzeniu tylu kolumn ile jest unikalnych wartości dla kategorii i wpisanie 0 lub 1 dla każdego rekordu danych [18] .

Współliniowość cech Aby znaleść korelacje współliniowości należy szukać liniowej zależności pomiędzy danymi, najłatwiej zauważyć to tworząc wykresy z danych testowych dla każdej pary [19] .

[TODO] Wwykresy dla cech

[18] Introduction to Data Preprocessing in Machine Learning Beginners Guide for Data Preprocessing Dhairya Kumar
[19] Introduction to Data Preprocessing in Machine Learning Beginners Guide for Data Preprocessing Dhairya Kumar

Opis praktycznej częsci projektu

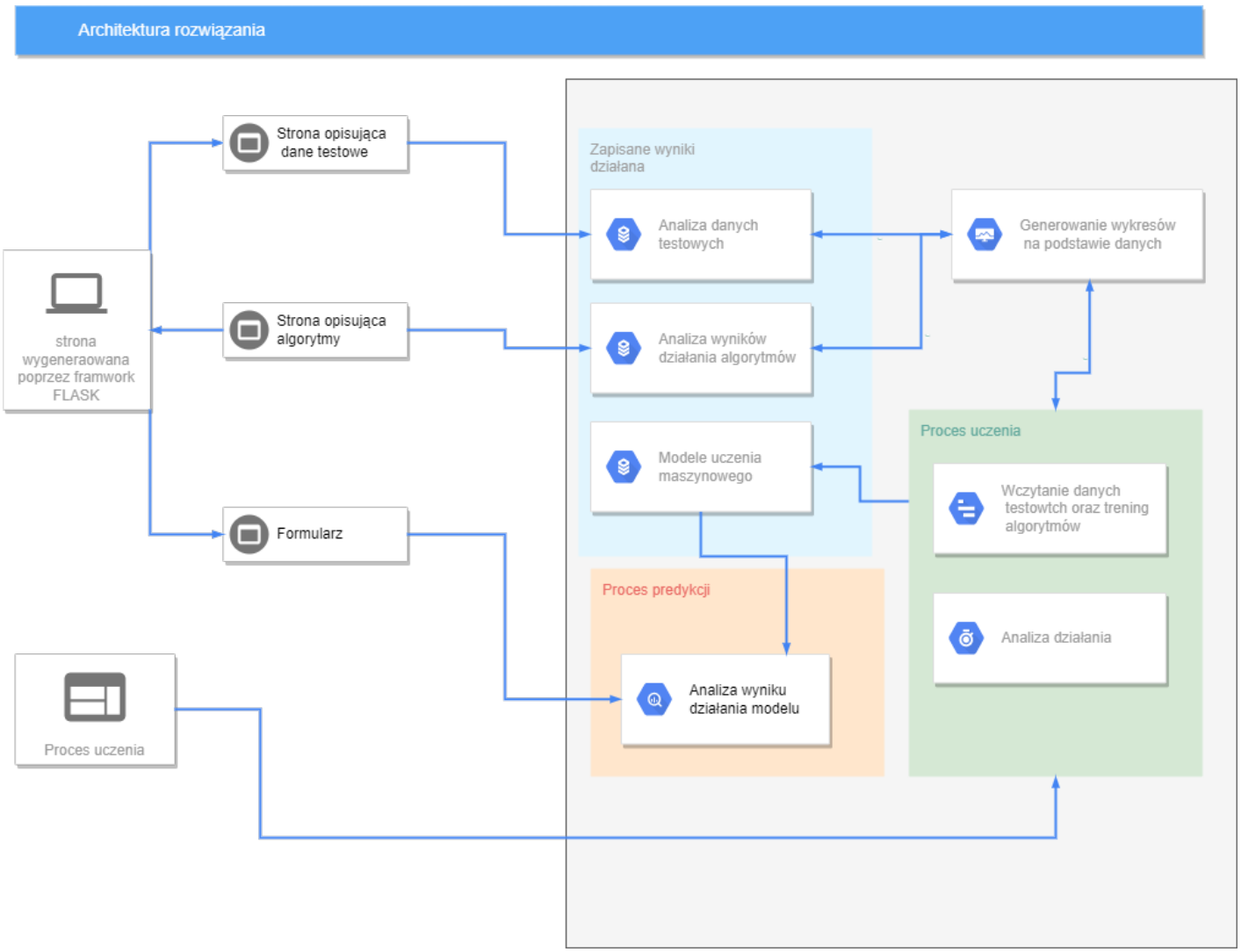
Moduły projektu:

- Config - zawiera statyczne zasoby oraz konfigurację logowania projektu
- Data - moduł odpowiada za wczytywanie i obróbkę danych testowych, oraz zawiera definicje objektów wykorzystywanych przy uczeniu oraz zapisu modelu
- Management:
  - PlotGeneration - moduł odpowiedzialny za prezentację wyników w postaci wykresów porównujących algorytmy oraz odpowiedzi na zadany problem
  - Prediction :
    - RF - implepmentacja treningu algorytmu Lasów losowych
    - KNN - implepmentacja treningu algorytmu K-najbliższych sąsiadów
    - SVM - implepmentacja treningu algorytmu Maszyny wektorów nośnych
- Static - forlder z grafikami, plikami stylów, skryptami javascript i jQuerry
- Templates - folder z stronami html wykorzystującymi dyrektywy Flask

Projekt posiada dwa tryby pracy :

- tryb nauczania na podstawie danych testowych - machine learning z wykorzystaniem 3 algorytmów (*Run_Learning_Proces.xml*)
- tryb aplikacji web - wykorzystanie Flask do prezentacji i wykorzystania utworzonych modeli (*Run_Web_Application.xml*)

Poniżej przedstawiono plan działania:



[todo] opisać główne metody

Narzędzia i biblioteki zastosowane w pojekcie

Praktyczna część pracy napisana została w języku Python z wykorzystaniem *scikit-learn*, obsługującym wiele algorytmów maszynowego uczenia się w tym uczenia nadzorowanego i docelowo wybranych algorytmów przedstawionych w teoretycznej części pracy.

Biblioteka opiera się o *Numpy* oraz *Scipy*, daje zestaw narzędzi do obliczeń na macierzach, wektorach oraz umożliwiający metody numeryczne takie jak całkowanie, różniczkowanie i temu podobne [20]. W rezultacie można za jej pomoca wykonać elementy procesu nauczania algorytmu, takie jak: przetwarzanie wstępne, redukcja wymiarowości, klasyfikacja, regresja.

[20]@article{scikit-learn, title={Scikit-learn: Machine Learning in {P}ython}, author={Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.}, journal={Journal of Machine Learning Research}, volume={12}, pages={2825–2830}, year={2011}}

[21]

Do przygotowania danych wykorzystano zestaw narzędzi *Pandas*, ułatwiający tworzenie struktur danych i ich analizę.

W celu wizualizacji wyników w postaci wykresów zastosowano, opartą na *Matplotlib*, bibliotekę *Seaborn* powszechnie stosowaną do rysowania estetycznej grfiki statystycznej.

Część prezentacyjna czyli możliwość wprowadzenia danych w formularzu na stronie i weryfikacja wyniku dla wyuczonych już modeli wykorzystuje bibliotkę *Flask*. Framework Flask ułatwia pisanie aplikacji internetowch ponieważ jest rozwiązaniem które daje duży zakres dowolności oraz możliwości. Flask sam z siebie nie definiuje warstwy bazy danych czy formularzy, pozwala za to na obsługę rozszerzeń które ubogacają aplikację o wybraną funkcjonalność. [22]

Przekazywanie obiektów o bardziej skomplikowanej budowie i ich *serializacja* oraz *deserializacja* do formatu JSON wykonane są za pomocą biblioteki *jsonpickle*, a zapis modeli wykonano za pomocą *joblib* która zapewnia obsługę obiektów Pythona i jest zoptymalizowana pod kątem pracy na dużych tablicach Numpy. [23]

Biblioteki w większości posiadają otwarty kod źródłowy , głównie napisany w języku Python [24].

Trening algorytmu

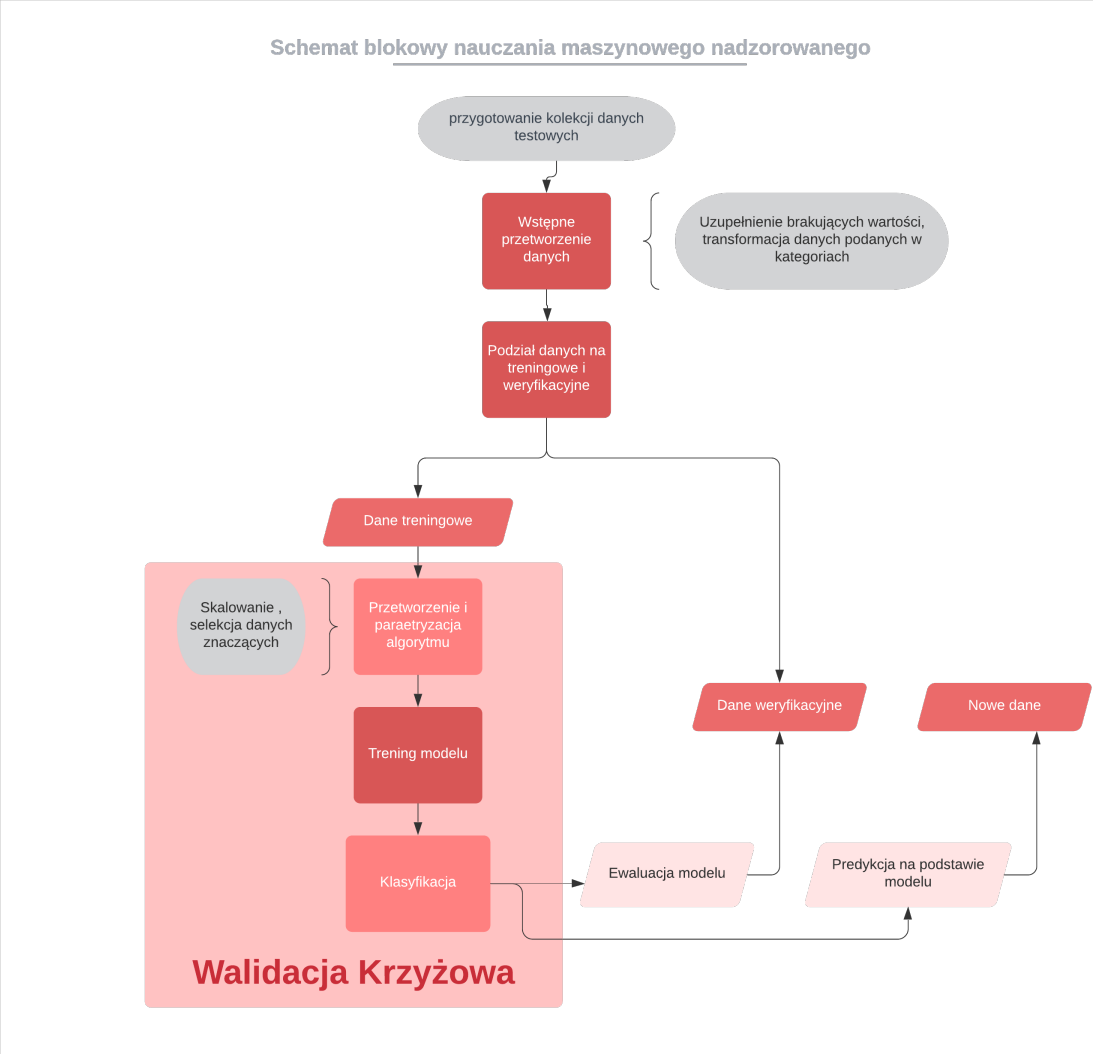
Zgodnie z poniższym schematem po przeprocesowaniu wejściowego zbioru danych, należy go podzielić na dane treingowe oraz ewaluacyjne. Powszechnie stosowana K krzyżowa walidacja umożliwia maksymalne wykorzystanie dostarczonego wejścia do dostrajania parametrów modelu, ponieważ optymalizacja hiperparametrów połączone z ciągłą weryfikacją poprawności to sedno treningu.

[21] Podejście porównawcze do algorytmów uczenia się maszynowego ,Samrudhi Rajendra Kaware , Vinod Subhasharao Wande

[22] @book{grinberg2018flask, title={Flask web development: developing web applications with python}, author={Grinberg, Miguel}, year={2018}, publisher={" O'Reilly Media, Inc."}}

[23] Podejście porównawcze do algorytmów uczenia się maszynowego ,Samrudhi Rajendra Kaware , Vinod Subhasharao Wande

[24] Podejście porównawcze do algorytmów uczenia się maszynowego ,Samrudhi Rajendra Kaware , Vinod Subhasharao Wande



K-krotna walidacja krzyżowa (ang. *K-fold Cross Validation* ,KCV) - metoda weryfikacji działająca poprzez podział zbioru danych na k podzbiorów z których każdy przynajmniej raz jest zbiorem oceniającym wydajność , zaznaczając że K musi być równe lub mniejsze niż liczba elementów w zbiorze [25] .

Kluczowym elementem jest ewaluacja która odbywa się na końcu każdej z k-1 iteracji w celu dostosowania parametrów, po osiągnięciu wymaganych lub ustalonych wartości dokładności modelu lub weryfikacji wszystkich możliwych opcji i znalezienie najlepszego modelu można go wykorzystać do weryfikacji na danych spoza zestawu testowego.

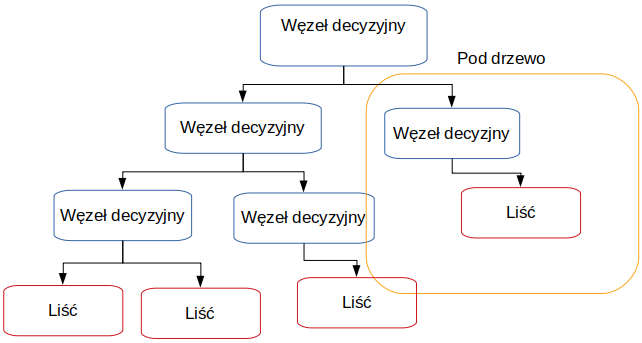
Wybrane algorytmy uczenia maszynowego nadzorowanego

Losowe lasy decyzyjne

Drzewa decyzyjne (ang. *decisions trees*) są uznawane za najprostyszy i najbliższy ludzkiemu zrozumieniu algorytm uczenia, który swoją nazwę zawdzięcza graficznej reprezentacji w postaci drzewa. Każdy węzeł oznacza atrybut, na podstawie którego następuje rozróżnienie. W modelu kluczowa jest kolejnośc cech, które występują po sobie ponieważ determinuje to otrzymany rezultat [26] .

[25] The 'K' in K-fold Cross Validation Authors: D. Anguita,L. Ghelardoni,A. Ghio,ONETO, LUCA,S. Ridella oraz Mastering Machine Learning Algorithms: Expert techniques to implement popular machine learning algorithms and fine-tune your models Giuseppe Bonaccorso

[26] Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8



Prawie każdy algorytm uczenia maszynowego nadzorowanego można podzielić na dwa etapy. W pierwszym opracowywany jest wzorzec, na którym bazuję późniejsza predykcja. Etap nauki dla drzewa decyzyjnego polega na typowaniu atrybutów, które staja się węzłami decyzyjnymi, dzielącymi rekordy na dwa mniejsze zestawy i tak aż nie ma możliwości dalszego podziału.

Na metodologie drzew decyzyjnych oparta jest dokładniejsza forma nauczania nadzorowanego: *losowe lasy decyzyjne.*

Losowe lasy decyzyjne (ang. *random decision forests*) to technika polegająca na połączeniu wielu drzew decyzyjnych w celu uniknięcia problemu z *nadmiernym dopasowaniem* do treningowego zestawu danych na którym został przeszkolony.

Utworzony szablon aby poprawnie działać na danych testowych i służacych weryfikacji, nie może stać się charakterystycznym przypadkiem rozwiązującym przypadek testowy [27]. W tym celu dla loswoych lasów decyzyjnych najpierw stosuję się **agregację bootstrap'ową**. Z treningowego zestawu danch losuję się, z możliwymi powtorzeniami, wiersze danych dla których trenowany będzie model. Jako rezutat brana jest większość lub średnia wartości uzyskanych wyników dla poszczególnych drzew dezycyjnch. Dodatkowo dla drzew decyzyjnych w lasach losowych, atrybuty odpowiadające za kategoryzację są wybierane z wylosowanego podzbioru.[28]

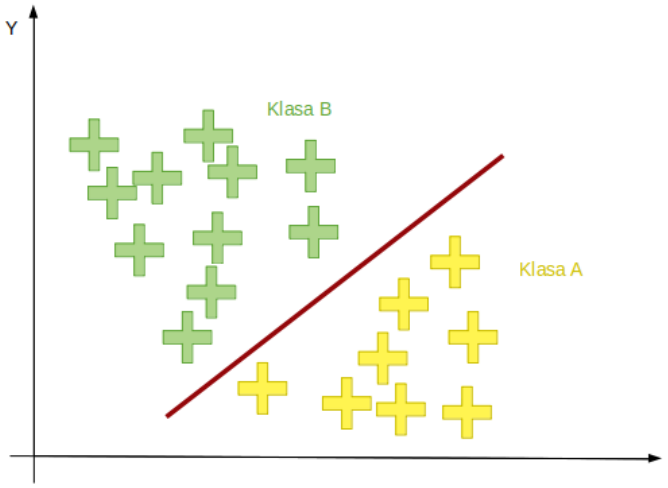
Wśrod zalet lasów losowych nalezy wyróżnić iż potrafią one trafnie wykalkulować brakujace wartości cech. Idealnie znajdują zastosowanie dla realnych danych, których zasadniczym problemem jest ich niekompletność.
Dane medyczne posiadają szeroką wariację zmiennych z dużym prawdopodobieństwem wybrakownia, zastosowanie do nich lasów decyzyjnych ma potencjał na pozytywne rezulataty.

Maszyna wektorów nośnych

Metoda wektorów nośnych (ang. *support vector machines* , skr. ***SVM***) to algorytm uczenia maszynowego nadzorowanego, który każdy parametr z dostępnych cech dla danych wejściowych, traktuje jako punkt w przestrzeni. Na podstawie ułożenia punktów dzieli się je na 2 klasy. Graficznie jest to reprezentowane przez prostą dla której odległość między najbliższymi dwoma punktami dla wektorów jest możliwie największa.

[27] Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

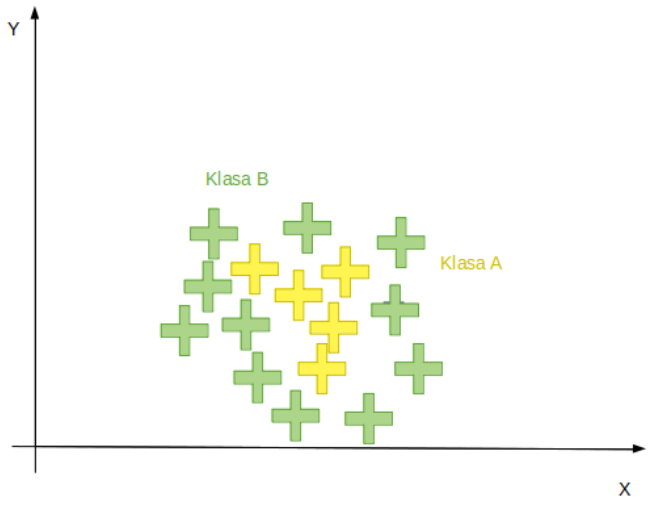
[28] Breiman, L. (2001), Random forests, Machine Learning 45: 5–32, FROM SINGLE TREES TO A RANDOM FOREST Tomasz Demski, StatSoft Polska Sp. z o. o



29

Taka prosta nazywana jest *prostą marginalną* i powstaje ona poprzez generowanie i selekcję tych prostych które rzetelnie szufladkują klasy danych [30].

Techinka ta gwarantuje precyzyjniejsze rezulatay niż drzewa decyzyjne, niestety dla dużych zbiorów danych czas trwania szkolenia znacznie się wydłuża oraz istnieją przypadki dla których podział jedną prostą jest niewykonalny, taki przypadek reprezentuje rozkład na schemacie nr. 2.



31

Z powyższego schematu widać że prosta marginalna ma zastosowanie w przypadku dwóch wymiarów, dla większej ilości stosowane jest przekształcenie do innego systemu współrzędnych i szukanie hiperpłaszczyzny brzegowej dzielącej tak samo jak prosta punkty w przestrzeni na dwa zbiory.[32]

Wyszukiwanie podziału

Idea działania maszyny wektorów nośnych opiera się na wyznaczenia minimalnej wartości wektora wag oraz przesunięcia (ang. *bias*) który geometrycznie opisuje współrzędne hiperpłaszczyzny.

[29] Na podstwie materiałów opublikowanych na https://www.datacamp.com

[30] Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

[31] Na podstwie materiałów opublikowanych na https://www.datacamp.com

[32] @article {HUANG41,author = {HUANG, SHUJUN and CAI, NIANGUANG and PACHECO, PEDRO PENZUTI and NARRANDES, SHAVIRA and WANG, YANG and XU, WAYNE}, title = {Applications of Support Vector Machine (SVM) Learning in Cancer Genomics}, volume = {15}, number = {1}, pages = {41–51}, year = {2018}, publisher = {International Institute of Anticancer Research}, issn = {1109-6535}, URL = {https://cgp.iiarjournals.org/content/15/1/41}, eprint = {https://cgp.iiarjournals.org/content/15/1/41.full.pdf}, journal = {Cancer Genomics & Proteomics}}

[Schemat 13](img/16svm_wzor2.png " svm wzor ") [33]

K najbliższych sąsiadów

K najbliższych sąsiadów (ang. *k nearest neighbours*, skr. ***KNN***) to algorytm uczenia maszynowego nadzorowanego operający swoje estymacje dla konkretnego przypadku danch na wartościach jego K najbliższych sąsiadów(punktów) liczonych min. dla przestrzeni Euklidesowej [34]. Do wyznaczenia odległości w metryce Euklidesowej stosowany jest wzór:

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad 35$$

popularne są również przestrzenie Manhattan:

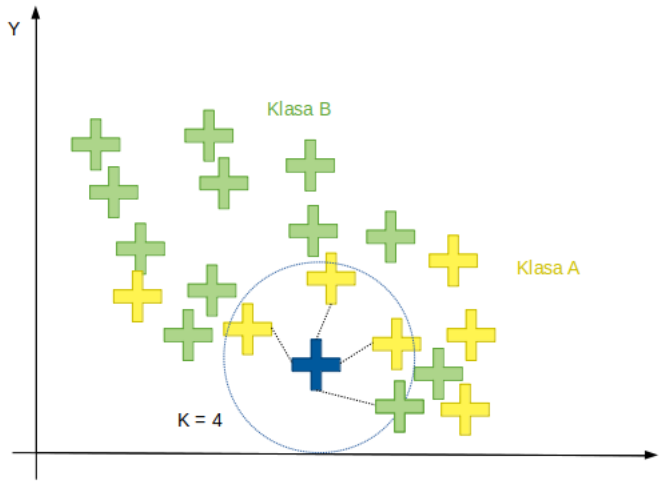
$$d_{(x,y)} = \sum_{i=1}^{n}|(x_i - y_i)| \quad 36$$

oraz Mińkowskiego:

$$d = \left(\sum_{i=1}^{m}|u_i - v_i|^p\right)^{1/p} \quad 37$$

Atrybut który nastraja proces uczenia się modelu i ma na niego największy wpływ określany jest jako hiperparametr. Dla KNN jest to liczba sąsiadów, im większa ilość jednostek mających wpływ, tym wierniejsze będą wyniki. Potęguje się wtedy niestety złożoność czasowa algorytmu, znacząco już większa od przedstawionych powyżej innych algoryrtmów.[38]

W celu przewidzenia wartości dla nowych danych, należy odnaleść K najbliższych punktów wyliczając odległości, a następnie przpisać odpowiedź implikowaną przez większość sąsiadów. Dla wartości K równej jeden, metoda ta nazywana jest algorytmem najbliższego sąsiada. K może przyjmować maksymalnie wartości do rozmiaru zbioru cech, jednak im bardziej są to zbliżone wartości tym bardziej wzrasta ryzyko nadmiernego dopasowania do modelu testowanego.



39

[33] Maszyna Wektorów Nośnych,Anna Pielowska

[34] Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

[35] Comparison of A*, Euclidean and Manhattan distance using Influence Map in Ms. Pac-Man aut.:Sudip Karki,Hari Sagar Ranjitkar,Faculty of Computing Blekinge Institute of Technology

[36] Comparison of A*, Euclidean and Manhattan distance using Influence Map in Ms. Pac-Man aut.:Sudip Karki,Hari Sagar Ranjitkar,Faculty of Computing Blekinge Institute of Technology

[37] The Minkowski approach for choosing the distance metric in geographically weighted regression Binbin Lua , Martin Charltonb , Chris Brunsdon and Paul Harrisc , School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland; Sustainable Soils and Grassland Systems, Rothamsted Research, North Wyke, Okehampton, Devon, UK

[38] Data Science from Scratch:First Principles with Python, Joel Grus, R.11,str140, Thoughtful Machine Learning with Python A Test-Driven Approach autor :Kirk Matthew r.1 str.8

[39] Na podstwie materiałów opublikowanych na https://www.datacamp.com

Dla lekarza wartością dodatnią jest wykrycie zależności które decyzują o uznaniu lub zaprzeczeniu występowania choroby. Zastosowanie algorytmu KNN może nie tylko zakwalifikować osoby chorujące na serce, ale również ułatwić swoją graficzną reprezentacją wpływ cech na ostateczny osąd próbki.

Komparacja działania modeli

impementacja z skleran która powstała w oparciu o dokumntacje sklearn

W tym podrozdziale zamieszczone zostały wyniki oraz wykresy wygenerowane podczas treningu i weryfikacji danych testowych, dla każdego algorytmu wykonano k-krotną walidację z wykorzystaniem:

GridSearchCV

do dostrojenia parametrów oraz znalezienia najlepszego modelu, dla każdego algorytmy zapróbkowano wszytkie dostępne dla danego modelu regresji parametry.

Wyczerpujące wyszukiwanie określonych wartości parametrów dla estymatora.

Ważni członkowie są sprawni, przewiduj.

GridSearchCV implementuje metodę „dopasowania" i „punktacji". Implementuje również „score_samples", „predict", „predict_proba", „decision_function", „transform" i „inverse_transform", jeśli są zaimplementowane w używanym estymatorze.

Parametry estymatora używanego do zastosowania tych metod są optymalizowane przez krzyżowo zweryfikowane wyszukiwanie w siatce parametrów.

Hiperparametry to parametry, których nie można nauczyć się bezpośrednio w estymatorach. W scikit-learn są one przekazywane jako argumenty do konstruktora klas estymatorów. Typowe przykłady to C, kernel i gamma dla Support Vector Classifier, alfa dla Lasso itp.

Możliwe i zalecane jest przeszukanie przestrzeni hiperparametrów w celu uzyskania najlepszego wyniku walidacji krzyżowej.

W ten sposób można zoptymalizować dowolny parametr podany podczas konstruowania estymatora. W szczególności, aby znaleźć nazwy i aktualne wartości wszystkich parametrów dla danego estymatora, użyj:

estymator.get_params() Wyszukiwanie składa się z:

estymator (regresor lub klasyfikator, taki jak sklearn.svm.SVC());

przestrzeń parametrów;

metoda wyszukiwania lub próbkowania kandydatów;

schemat walidacji krzyżowej; oraz

funkcja punktacji.

W scikit-learn dostępne są dwa ogólne podejścia do wyszukiwania parametrów: dla podanych wartości GridSearchCV w sposób wyczerpujący uwzględnia wszystkie kombinacje parametrów, podczas gdy RandomizedSearchCV może próbkować określoną liczbę kandydatów z przestrzeni parametrów o określonym rozkładzie. Oba te narzędzia mają kolejne odpowiedniki HalvingGridSearchCV i HalvingRandomSearchCV, które mogą znacznie szybciej znaleźć dobrą kombinację parametrów.

Po opisaniu tych narzędzi szczegółowo opisujemy najlepsze praktyki mające zastosowanie do tych podejść. Niektóre modele pozwalają na wyspecjalizowane, wydajne strategie wyszukiwania parametrów, opisane w Alternatywach do wyszukiwania parametrów metodą brute force.

Należy zauważyć, że często mały podzbiór tych parametrów może mieć duży wpływ na wydajność predykcyjną lub obliczeniową modelu, podczas gdy inne można pozostawić z wartościami domyślnymi. Zaleca się zapoznanie się z dokumentacją klasy estymatora, aby lepiej zrozumieć ich oczekiwane zachowanie, prawdopodobnie poprzez przeczytanie załączonych odnośników do literatury.

3.2.1. Wyczerpujące wyszukiwanie w siatce Wyszukiwanie siatki zapewniane przez GridSearchCV w sposób wyczerpujący generuje kandydatów z siatki wartości parametrów określonych za pomocą parametru param_grid. Na przykład następujący param_grid:

param_grid = [{‘C’: [1, 10, 100, 1000], ‘kernel’: [‘linear’]}, {‘C’: [1, 10, 100, 1000], ‘gamma’: [0.001, 0.0001], ‘jądro’: [‘rbf’]},] określa, że należy zbadać dwie siatki: jedną z jądrem liniowym i wartościami C w [1, 10, 100, 1000], a drugą z jądrem RBF i iloczynem krzyżowym wartości C w zakresie [1, 10 , 100, 1000] i wartości gamma w [0,001, 0,0001].

Instancja GridSearchCV implementuje zwykły interfejs API estymatora: podczas „dopasowywania” go do zbioru danych oceniane są wszystkie możliwe kombinacje wartości parametrów i zachowywana jest najlepsza kombinacja.

##estymator obiekt estymatora Zakłada się, że jest to implementacja interfejsu estymatora scikit-learn. Albo estymator musi podać funkcję punktacji, albo punktacja musi zostać przekazana.

param_griddict lub lista słowników Słownik z nazwami parametrów (str) jako kluczami i listami ustawień parametrów do wypróbowania jako wartości lub listą takich słowników, w którym to przypadku eksplorowane są siatki zawarte w każdym słowniku na liście. Umożliwia to przeszukiwanie dowolnej sekwencji ustawień parametrów.

scoringstr, wywoływalne, lista, krotka lub dyktowanie, domyślnie=Brak Strategia oceny wydajności modelu poddanego walidacji krzyżowej na zbiorze testowym.

Jeżeli punktacja reprezentuje pojedynczą punktację, można użyć:

pojedynczy ciąg (patrz Parametr scoringowy: definiowanie reguł oceny modelu);

wywoływalna (zobacz Definiowanie strategii scoringowej na podstawie funkcji metryki), która zwraca pojedynczą wartość.

Jeśli punktacja reprezentuje wiele punktów, można użyć:

lista lub krotka unikalnych ciągów;

wywoływalny zwracający słownik, w którym klucze są nazwami metryk, a wartości są wynikami metryk;

słownik z nazwami metryk jako kluczami i wywoływalnymi wartościami.

Zobacz na przykład Określanie wielu metryk do oceny.

n_jobsint, domyślnie=Brak Liczba zadań do równoległego uruchomienia. Brak oznacza 1, chyba że w kontekście joblib.parallel_backend. -1 oznacza użycie wszystkich procesorów. Zobacz Słowniczek po więcej szczegółów.

Zmieniono w wersji v0.20: domyślna wartość n_jobs zmieniona z 1 na Brak

refitbool, str lub callable, default=True Dopasuj estymator, korzystając z najlepszych znalezionych parametrów w całym zbiorze danych.

W przypadku oceny wielu metryk musi to być str oznaczający punktację, który zostałby użyty do znalezienia najlepszych parametrów do ponownego dopasowania estymatora na końcu.

Tam, gdzie przy wyborze najlepszego estymatora istnieją względy inne niż maksymalny wynik, refit można ustawić na funkcję, która zwraca wybrany najlepszy_indeks_ podany cv_results_. W takim przypadku best_estimator_ i best_params_ zostaną ustawione zgodnie ze zwróconym best_index_, podczas gdy atrybut best_score_ nie będzie dostępny.

Dopasowany estymator jest udostępniany w atrybucie best_estimator_ i umożliwia użycie predykcji bezpośrednio w tym wystąpieniu GridSearchCV.

Również w przypadku oceny wielu metryk atrybuty best_index_, best_score_ i best_params_ będą dostępne tylko wtedy, gdy ustawiony jest remont, a wszystkie z nich zostaną określone z uwzględnieniem tego konkretnego scoringowca.

Zobacz parametr scoring, aby dowiedzieć się więcej o ocenie wielu metryk.

Zmieniono w wersji 0.20: Dodano obsługę callable.

cvint, generator walidacji krzyżowej lub iterowalny, domyślnie=Brak Określa strategię podziału z walidacją krzyżową. Możliwe dane wejściowe dla CV to:

Brak, aby użyć domyślnej pięciokrotnej weryfikacji krzyżowej,

liczba całkowita, aby określić liczbę fałd w (Stratified)KFold,

rozdzielacz CV,

Iterowalny plon (pociąg, test) dzieli się na tablice indeksów.

W przypadku danych wejściowych typu liczba całkowita/brak, jeśli estymator jest klasyfikatorem, a y jest binarne lub wieloklasowe, używany jest StratifiedKFold. We wszystkich innych przypadkach używany jest KFold. Te splittery są tworzone z shuffle=False, więc podziały będą takie same we wszystkich wywołaniach.

Zapoznaj się z podręcznikiem użytkownika, aby zapoznać się z różnymi strategiami walidacji krzyżowej, których można tu użyć.

Zmieniono w wersji 0.22: domyślna wartość cv, jeśli Brak zmieniła się z 3-krotnej na 5-krotną.

verboseint Kontroluje szczegółowość: im wyższa, tym więcej wiadomości.

1 : wyświetlany jest czas obliczeń dla każdego fałdu i potencjalnego parametru;

2 : wyświetlany jest również wynik;

3 : indeksy parametrów fałd i kandydatów są również wyświetlane wraz z czasem rozpoczęcia obliczeń.

pre_dispatchint lub str, default='2*n_jobs' Kontroluje liczbę zadań, które są wysyłane podczas wykonywania równoległego. Zmniejszenie tej liczby może być przydatne, aby uniknąć eksplozji zużycia pamięci, gdy wysyłanych jest więcej zadań, niż może przetworzyć procesor. Ten parametr może być:

Brak, w takim przypadku wszystkie miejsca pracy są natychmiast tworzone i odradzane. Użyj tego do lekkich i szybko działających zadań, aby uniknąć opóźnień spowodowanych pojawianiem się zadań na żądanie

Int, podający dokładną liczbę wszystkich miejsc pracy, które się odradzają

A str, dające wyrażenie w funkcji n_jobs, jak w „2*n_jobs"

error_score'podniesienie' lub numeryczne, domyślnie=np.nan Wartość do przypisania do wyniku, jeśli wystąpi błąd w dopasowaniu estymatora. Jeśli ustawione na „podnieś", błąd jest zgłaszany. Jeśli zostanie podana wartość liczbowa, zostanie zgłoszone FitFailedWarning. Ten parametr nie ma wpływu na etap naprawy, który zawsze spowoduje zwiększenie błędu.

return_train_scorebool, domyślnie = Fałsz Jeśli False, atrybut cv_results_ nie będzie zawierał wyników szkolenia. Obliczanie wyników treningowych służy do uzyskiwania wglądu w to, jak różne ustawienia parametrów wpływają na kompromis polegający na przesunięciu/niedopasowaniu. Jednak obliczanie wyników na zbiorze uczącym może być kosztowne obliczeniowo i nie jest ściśle wymagane do wyboru parametrów, które zapewniają najlepszą wydajność uogólniania.

Nowość w wersji 0.19. []

##todo liczenie błędów macież pomysłek

Resultaty wnioski: Losowe lasy decyzyjne

###OCENA PODELI ORAZ UŻYTYCH PARAMETRÓW -OCENA SZYBKOŚCI WYKONANIA -OCENA ZALEŻNIE OD UZUPELNIANIA DANYCH -OCENA ZALEŻNIE OD DOBRANEJ PARAMERYZACJI : - które parametry mają i wpływ i dlaczego: - ZALEŻNIE OD METRYKI(SHORT OPIS METRYK)

Rezultaty wnioski: Metoda wektorów nośnych

###OCENA PODELI ORAZ UŻYTYCH PARAMETRÓW -OCENA SZYBKOŚCI WYKONANIA -OCENA ZALEŻNIE OD UZUPELNIANIA DANYCH -OCENA ZALEŻNIE OD DOBRANEJ PARAMERYZACJI : - które parametry mają i wpływ i dlaczego: - ZALEŻNIE OD METRYKI(SHORT OPIS METRYK)

Rezultaty wnioski: K najbliższych sąsiadów

###OCENA PODELI ORAZ UŻYTYCH PARAMETRÓW -OCENA SZYBKOŚCI WYKONANIA -OCENA ZALEŻNIE OD UZUPELNIANIA DANYCH -OCENA ZALEŻNIE OD DOBRANEJ PARAMERYZACJI : - które parametry mają i wpływ i dlaczego: - ZALEŻNIE OD METRYKI(SHORT OPIS METRYK)

Plusy Faza uczenia klasyfikacji K-najbliższego sąsiada jest znacznie szybsza w porównaniu z innymi algorytmami klasyfikacji. Nie ma potrzeby uczenia modelu do uogólniania, dlatego KNN jest znany jako prosty algorytm uczenia oparty na instancjach. KNN może być przydatny w przypadku danych nieliniowych. Może być używany z problemem regresji. Wartość wyjściowa obiektu jest obliczana przez średnią k wartości najbliższych sąsiadów.

Cons Faza testowania klasyfikacji najbliższych sąsiadów K jest wolniejsza i bardziej kosztowna pod względem czasu i pamięci. Wymaga dużej pamięci do przechowywania całego zestawu danych treningowych do przewidywania. KNN wymaga skalowania danych, ponieważ KNN wykorzystuje odległość euklidesową między dwoma punktami danych, aby znaleźć najbliższych sąsiadów. Odległość euklidesowa jest wrażliwa na wielkości. Obiekty o dużych jasnościach będą miały większą wagę niż obiekty o niskich jasnościach. KNN nie nadaje się również do dużych danych wymiarowych.

Jak ulepszyć KNN? Aby uzyskać lepsze wyniki, zdecydowanie zaleca się normalizację danych w tej samej skali. Ogólnie rzecz biorąc, rozważany zakres normalizacji między 0 a 1. KNN nie jest odpowiedni dla danych wielkowymiarowych. W takich przypadkach wymiar musi się zmniejszyć, aby poprawić wydajność. Również obsługa brakujących wartości pomoże nam poprawić wyniki.

Porównianie całościowe algorytmów : złożoność czasowa , dokładność , złożoność implementacyjna , wpływ danych wykorzytywanych w modelu

porównanie z innymi pracami ktore robią klasyfikację rozwiązują problem jakiej metody użytli i jaki jest wynik ewalacji -> metody w porówananiu dają konukrencyjne wyniki hipotezy dlaczego tak się dzieje

Podsumowanie i opisanie wpływu danych na model

porównanie do dnych statystycznych

todo variants of user data preparatrio

```
        ## preparation all -> all test
        ## preparation best for best
        ## best from other to best in another  -> result and reasons for data anlayse
        ## fast not best - why is it faster
        ##
# todo prediction
# todo percentage na true false
```

problem multiklasyfikacji - problem regresji kategrycznej - zwykła regresja , mierzyć będe metoda prównania - tzrea było wprowadzić reguły do float na int -> inne metody do liczenia błędów na dzień dobry widzimy nie dokładność ze wględu na klasyfiakcję po przecinku regresja kategoryczna -> rzutowanie przedziału wartości na wartość graniczną

Zestawienie efektywności działania algorytmów

Konfrontacja technik ucznia maszynowego zaleznie od zestawu danch będzie dawała odmienne wyniki ze względu na ich predyspozycje do zajmowania się odpowiednimi zbiorami danych.

Potencjał algorytmów dla niewielkiego kompletu danych zawierającego wartości wybrakowane zostanie omówony w późniejszych rozdziałach pracy.

Zczynając od drzew decyzyjnych, można od razu stwierdzić ich niski potencjał. Istnieje zbyt duże prawdopodobieństwo dopasowania się do modelu treningowego, gdyż wspomniany zbiór dancyh wejściowych nie jest wystarczająco liczny. Dlatego w dalszej części pracy omówione zostaną lasy decyzyjne.

Większej dokładności można się spodziewać po metodzie wektorów nośnych, ale jego złożoność czasowa oraz pamięciowa mogą zaniżyc jego ogólną klasyfikację.

Wskaźniki wydajności

Określenie stopnia, w jakim skonstruowany model z powodzeniem realizuje wyznaczone zadanie należy do wskaźnika wydajności. Przykładem nieprawidłowego wyboru może być próba przewidzenia wystąpienia rzadkiej choroby u pacjenta i określenie głownym miernikiem *dokładność*. W takim scenariuszu klasyfikacja wszystkich pacjentów jako zdrowych , daje niewiele odbiegającą od perfekcji dokładność, a jednocześnie błędnie osądzać każde wystąpienie choroby.

Spis ilustracji

Spis tabel

Bibliografia

@article{scikit-learn, title={Scikit-learn: Machine Learning in {P}ython}, author={Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.}, journal={Journal of Machine Learning Research}, volume={12}, pages={2825–2830}, year={2011} }

- @article{http://www.mif.pg.gda.pl/homepages/kdz/BIGDATA/AniaPielowska.pdf}
- @article{https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/}
- @article{https://myservername.com/what-is-support-vector-machine-machine-learning}
- @article{https://scikit-learn.org/stable/modules/svm.html}
- @article{https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/}
- @article{https://scikit-learn.org/stable/modules/neighbors.html}
- @article{https://scikit-learn.org/stable/modules/naive_bayes.html}
- @article{https://scikit-learn.org/stable/modules/tree.html}
- @article{https://scikit-learn.org/stable/modules/feature_selection.html}
- @article{http://pages.cs.wisc.edu/~dpage/kuusisto.thesis.pdf}
- @article{http://www.bme.teiath.gr/medisp/pdfs/PhD_Glotsos_Dimitrios.pdf}
- @article{https://www.springboard.com/blog/how-to-become-a-machine-learning-engineer/}

- @article{http://www.diva-portal.org/smash/get/diva2:920202/FULLTEXT01.pdf}
- @article{https://www.techsparks.co.in/hot-topic-for-project-and-thesis-machine-learning/}
- @article{https://machinelearningmastery.com/k-fold-cross-validation/}
- @article{https://www.writemythesis.org/master-thesis-topics-in-machine-learning/}
- @article{http://mediatum.ub.tum.de/doc/1368117/47614.pdf}
- @article{https://pdfs.semanticscholar.org/0e06/561dbab0581feebe6638dc2671f94c9abf68.pdf}
- @article{https://www.cir.meduniwien.ac.at/assets/Uploads/Masterthesis-SeeboeckPhilipp-Version28-03-2015.pdf}
- @article{https://www.quora.com/Is-there-any-machine-learning-thesis-idea-in-health-care}
- @article{https://digitalcommons.odu.edu/cgi/viewcontent.cgi?referer=- @article{https://www.google.pl/&httpsredir=1&arti
- @article{https://www.mobt3ath.com/uplode/book/book-60163.pdf}
- @article{https://www.ilovephd.com/thesis-bank-machine-learning-2/}
- @article{https://www.digitalocean.com/community/tutorials/how-to-handle-plain-text-files-in-python-3}
- @article{https://machinelearningmastery.com/naive-bayes-for-machine-learning/}
- @article{https://machinelearningmastery.com/compare-machine-learning-algorithms-python-scikit-learn/}
- @article{https://machinelearningmastery.com/compare-machine-learning-algorithms-python-scikit-learn/}
- @article{https://elitedatascience.com/machine-learning-algorithms}
- @article{https://www.dataschool.io/comparing-supervised-learning-algorithms/}
- @article{https://medium.com/value-stream-design/online-machine-learning-515556ff72c5}
- @article{https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f}
- @article{https://www.kaggle.com/aldemuro/comparing-ml-algorithms-train-accuracy-90}
- @article{https://www.kaggle.com/aldemuro/comparing-ml-algorithms-train-accuracy-90}
- @article{https://machinelearningmastery.com/start-here/}
- @article{https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/}
- @article{https://blog.statsbot.co/machine-learning-algorithms-183cc73197c}
- @article{https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/}
- @article{https://scikit-learn.org/stable/modules/clustering.html}#overview-of-clustering-methods}
- @article{https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c}
- @article{https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c}
- @article{https://www.kaggle.com/cdabakoglu/heart-disease-classifications-machine-learning}
- @article{https://medium.com/@dskswu/machine-learning-with-a-heart-predicting-heart-disease-b2e9f24fee84}
- @article{https://pdfs.semanticscholar.org/d0a5/d4b8e8da3ee2a6bf8ac5d44196fb0365cf1c.pdf}
- @article{file:///home/szulce/Pobrane/Heart_Disease_Detection_by_Using_Machine_Learning_.p}df}
- @article{file:///home/szulce/Pobrane/jcm-08-01050.pdf}
- @article{http://www.cs.put.poznan.pl/alabijak/emd/12_Reprezentacja_wektorowa_slow.pdf}
- @article{https://www.hindawi.com/journals/misy/2018/3860146/}
- @article{https://pub.towardsai.net/3-different-approaches-for-train-test-splitting-of-a-pandas-dataframe-d5e544a5316}
-

@article{https://www.run.ai/guides/machine-learning-engineer/machine-learning-workflow/#:~:text=Machine%20learning%20workflows%20define%20wh

@article{https://www.dovepress.com/ensemble-approach-for-developing-a-smart-heart-disease-prediction-syst-peer-reviewed-fulltext-article-RRCC}

- @article{https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/}
-

@article{https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn?utm_source=adwords_pp
392016246653:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=1011615&gclid=Cj0KCQiA0eOPBhCGARIsA

- @article{https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/}
- @article{https://m.scirp.org/papers/88650}
- @article{https://link.springer.com/chapter/10.1007/978-3-540-24668-8_21}
- @article{https://erogol.com/machine-learning-work-flow-part-1/}
- @article{https://www.annualreviews.org/doi/pdf/10.1146/annurev-fluid-010719-060214}
- @article{https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94}
- @article{https://cloud.google.com/ai-platform/docs/ml-solutions-overview}

- @article{https://ai.ia.agh.edu.pl/__media/pl:dydaktyka:mbn:uczenie__maszynowe.pdf}
-

@article{https://www.researchgate.net/profile/Krzysztof-Krawiec/publication/235352247_Sieci__neuronowe__i__uczenie__maszynow
neuronowe-i-uczenie-maszynowe.pdf}

- @article{https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/pdf}
-

@article{https://www.statystyczny.pl/co-to-jest-machine-learning/#:~:text=Niekt%C3%B3rzy%20wspominaj%C4%85%20tu%20k

- @article{https://www.sciencedirect.com/science/article/pii/S1877050915024928}
- @article{https://machinelearningmastery.com/types-of-classification-in-machine-learning/}
- @article{https://data-flair.training/blogs/types-of-machine-learning-algorithms/}
- @article{https://ichi.pro/pl/co-to-jest-kodowanie-one-hot-i-jak-uzywac-funkcji-pandas-get-dummies-160729382340976}
- @article{https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5640485/}
- @article{https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/}
- @article{https://towardsdatascience.com/heart-disease-prediction-73468d630cfc}
- @article{https://www.sciencedirect.com/science/article/pii/S187705091630638X}
-

@article{https://www.ices.on.ca/Publications/Journal-Articles/2014/January/C Disease-Population-Risk-Tool-predictive-algorithm-for-assessing-CVD-risk}

@article{https://www.ctvnews.ca/health/test-your-risk-of-heart-disease-with-a-new-online-lifestyle-calculator-1.4030088}

- @article{https://nevonprojects.com/heart-disease-prediction-project/}
- @article{https://scikit-learn.org/stable/modules/neighbors.html}
- @article{https://searchenterpriseai.techtarget.com/definition/machine-learning-ML}
- @article{https://www.forcepoint.com/cyber-edu/machine-learning}
- @article{https://en.wikipedia.org/wiki/Supervised_learning}
- @article{https://www.techopedia.com/definition/8181/machine-learning}
- @article{https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/}
- @article{https://searchenterpriseai.techtarget.com/definition/supervised-learning}
- @article{https://deepai.org/machine-learning-glossary-and-terms/supervised-learning}
-

@article{https://towardsdatascience.com/what-are-supervised-and-unsupervised- learning-in-machine-learning-dc76bd67795d}

@article{https://towardsdatascience.com/what-are-supervised-and-unsupervised-learning-in-machine-learning-dc76bd67795d}

- @article{http://www.cs.ucr.edu/~mwile001/papers/thesis.pdf}
- @article{https://python-graph-gallery.com/111-custom-correlogram/}
- @article{https://python-graph-gallery.com/242-area-chart-and-faceting/}
- @article{https://web.stanford.edu/~hastie/Papers/ESLII.pdf}
- @article{https://www.sciencedirect.com/topics/computer-science/random-decision-forest}
- @article{https://flask.palletsprojects.com/en/1.1.x/tutorial/install/}
- @article{https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d}
- @article{https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html}
- @article{https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/}
- @article{https://dev.to/alod83/3-different-approaches-for-traintest-splitting-of-a-pandas-dataframe-31p0}
- @article{https://pub.towardsai.net/3-different-approaches-for-train-test-splitting-of-a-pandas-dataframe-d5e544a5316}
- @article{https://docs.python.org/3/library/itertools.html#itertools.zip_longest}
- @article{https://realpython.com/train-test-split-python-data/}
- @article{https://towardsdatascience.com/flask-and-chart-js-tutorial-i-d33e05fba845}

- @article{https://www.sciencedirect.com/science/article/pii/S2352914820300125 - pobrane jako pdfy}
- @article{https://docs.python.org/3/library/zipfile.html}
- @article{https://flask.palletsprojects.com/en/2.0.x/quickstart/}
- @article{https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/}
- @article{https://joblib.readthedocs.io/en/latest/}
- @article{https://www.kaggle.com/prmohanty/python-how-to-save-and-load-ml-models}
- @article{https://machinelearningmastery.com/machine-learning-in-python-step-by-step/}
-

@article{https://dobrebadania.pl/zmienna-dyskretna-ang-discrete-variable/#:~:text=Zmienna%20dyskretna%20to%20ka%C5%BC

- https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02
- https://www.ritchieng.com/machine-learning-efficiently-search-tuning-param/
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- https://www.vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv/
- http://jsonpickle.github.io/#jsonpickle-usage
- https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/
- https://ichi.pro/pl/jak-najlepiej-ocenic-model-klasyfikacji-51518447076743