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

Magdalena Szulc

2022-02-11

Contents

[Abstrakt]

[[TODO]]

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 ich podstawie wyszukiwane są zależności znaczące oraz budowany jest model do przewidywania wartości.

W przypadku danych dotyczących chorówb 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. Objawy są tym silniejsze im gorszy jest stan chorobowy pacjenta.

Dlatego prowadzone są intensywne prace nad zwiększeniem dostępności badań, które wspomogą diagnostykę kardiologiczną na jak najwcześniejszym etapie.

Powodem szukania dokładniejszych sposobów diagnozowania są również wysokie koszty leczenia generowane przez choroby układu krwionośnego. Według analityków firmy konsultingowej KPMG [1] w 2011 r. koszty diagnostyki i terapii chorób serca wyniosły ponad 15 miliardów polskich złotych.

Nadzieją jaka pokładana jest w machine learningu jest szybsza diagnostyka chorób ułatwiająca oraz przyśpieszająca proces ich leczenia. Zastosowanie uczenia maszynowego w medycynie, pozwala również na przetwarzanie dużych zasobów historycznych wyników medycznych, w celu poszerzenia zasobów informacji , głównie zależności przyczynowo skutkowych , które mogą zostać wykorzystane do diagnostyki lub leczenia.

Słowa kluczowe: uczenie maszynowe, uczenie nadzorowane

[1] 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

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:

- lasy decyzyjne (ang. *decisions-forests*)
- metoda wektorów nośnych (ang. *support vector machines*, SVM)
- k-najbliższych sąsiadów (ang. *k-neares neighbours*, KNN)

[TODO]słownictwo wykorzystywanego podczas pisania pracy tłumaczenia i wykorzystywane powszechnie w publikacjach naukowych

Repozytorium uczenia maszynowego UCI

Figure 1: Schemat 1

Sensem wykorzystania uczenia maszynowego jest przewidzenie lub klasyfikacja rzeczywistych wartości które można zastosować w innych dziedzinach. Im bardziej dokładne i rzeczywisty jest wsad do tworzenia modelu tym bardziej możliwe jest osiągnięcie lepszych efektów na końcu ścieżki uczeniia. W celu gromadzenia zaufanej bazy dostępnych zbiorów dataset'ów powstało repozytorium uczenia maszynowego UCI. Jak podaje strona informacyjna :

> [2]... było ono cytowane ponad 1000 razy, co czyni je jednym ze 100 najczęściej cytowanych „artykułów" w całej informatyce ...

[2]Źródło: https://archive.ics.uci.edu/ml/index.html

Wprowadzenie teorertyczne

Uczenie maszynowe (ang._machine learning_, ML) to dziedzina zajmująca się zestawem algorytmów, które analizując zbiory dancyh (zazwyczaj baerdzo obszerne) wystawiają predyckję na temat zadanego problemu. Uczenie maszynowe zależnie od sposobu *trenowania* algorytmu dzieli się na kategorie min. uczenie nadzorowane oraz uczenie bez nadzoru.

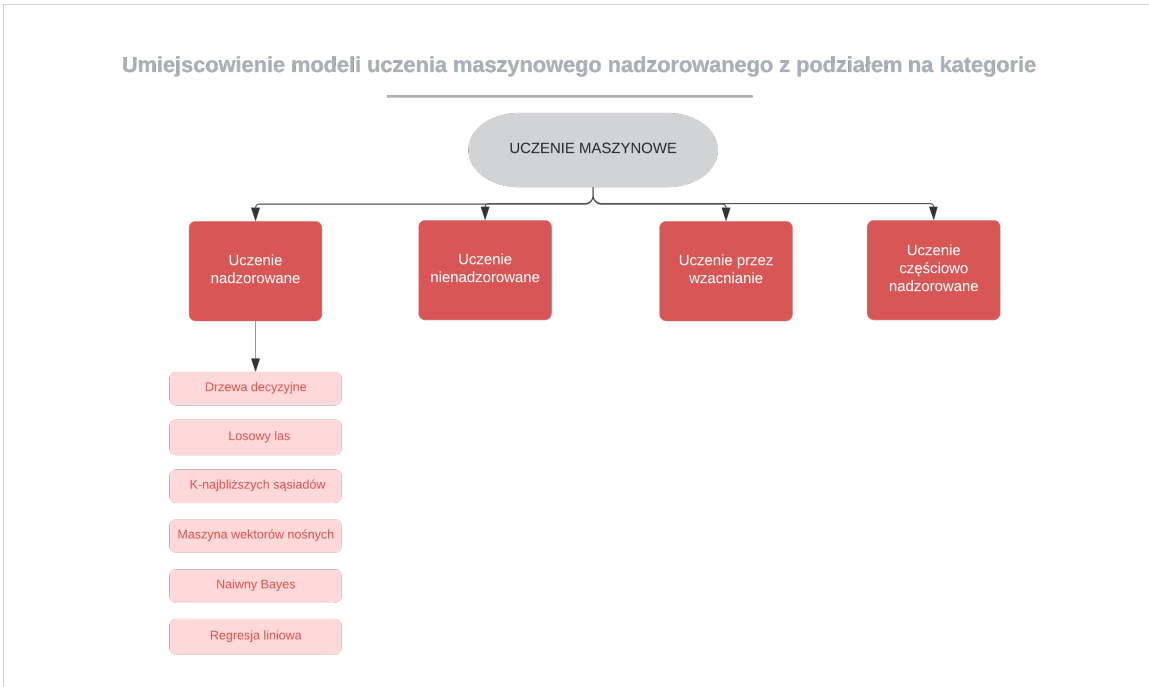

Figure 2: Schemat 1

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 już danych wejściowych. Ten typ uczenia świetnie nadaje się do rozwiązywania problemów z zakresu klasyfikacji. Nadzór polega na porównaniu rezultatów działania modelu z wynikami które są zawarte w danych wejściowych(*dane oznaczone*). Algorytm po osiągnięciu żądanej efektywnosci jest w stanie dokonać klasyfikacji przykładu dla którego nie posiada odpowiedzi. Sprawdza się to obecnie w rekomendacji produktów oraz diagnozie chorób.

Uczenie maszynowe bez nadzoru (ang. *unsupervised learning*) to klasa algroytmów uczenia maszynowego która głownie 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.

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

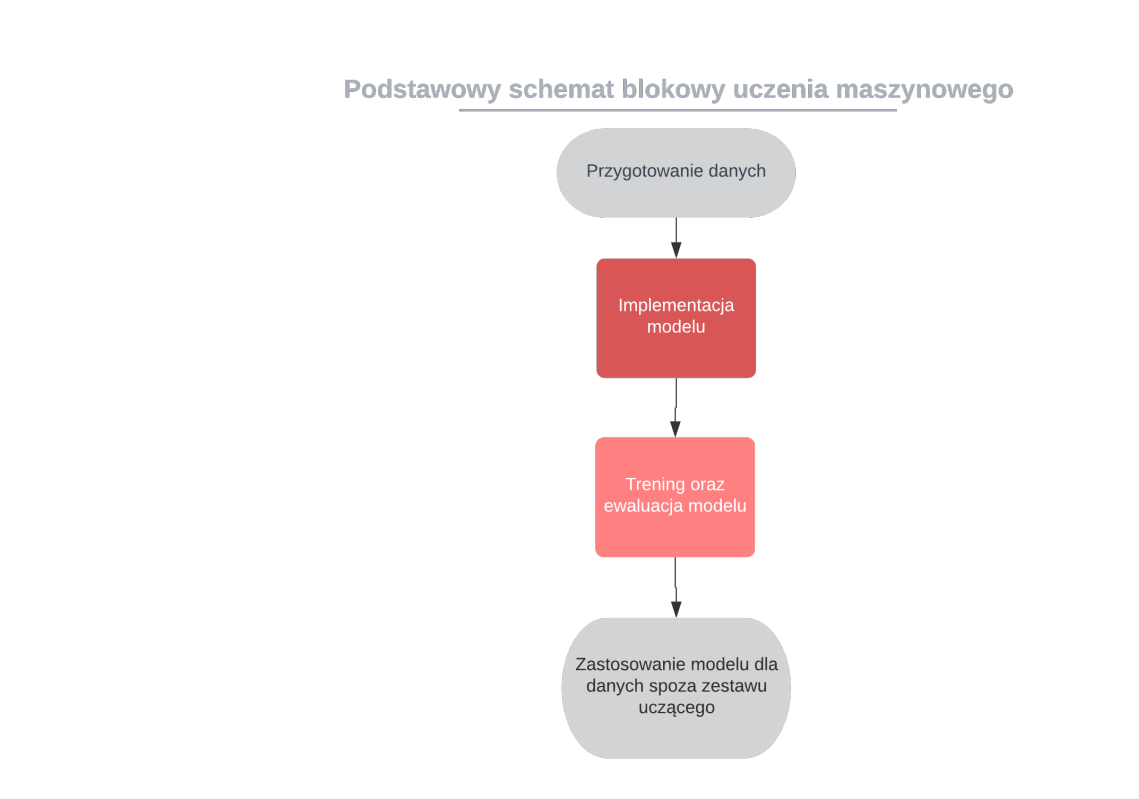

Figure 3: Schemat 1

Model Danych

Rozpoczęcie pracy nad budowaniem modelu dla algorytmów uczenia maszynowego w szeroko pojętym znaczeniu zawsze będzie zaczynało się od zebrania danych testowych, jest to czynnik determinujący wybór mmiędzy uczeniem z nadzorcą lub bez.

W przypadku danych testowych z repozytorium UCI , dane pochodziły z różnych lokalizacji , od tego zależeć mogą jakim badaniom poddani zostali pacjęci a co za tym idzie w jakich kolumnach tabelarycznego przedstawienia będą mieć uzupełnione bądź puste wartości. Scalenie ze sobą dataset'ów dostarcza większej wariacji. Jeżeli zestaw wejściowy został by ograniczony do jednej lokalizacji to cecha dla której nie uzupełniono wartości zostałaby z autoatu pominięta jako znacząca ze względu na brak danych. Po złączeniu można przeprowadzić szereg działań w celu sztucznego uzupełnienia pustych wartości bazując na wartościach które już istnieją.

Proces przetwarzania danych może skladać sie z wielu różnych kroków zależenie od typu, w uczeniu

nadzorowanym operującym na danych tekstowo-liczbowych poprawnym będzie zastosowanie schematu przedstawionego poniżej:

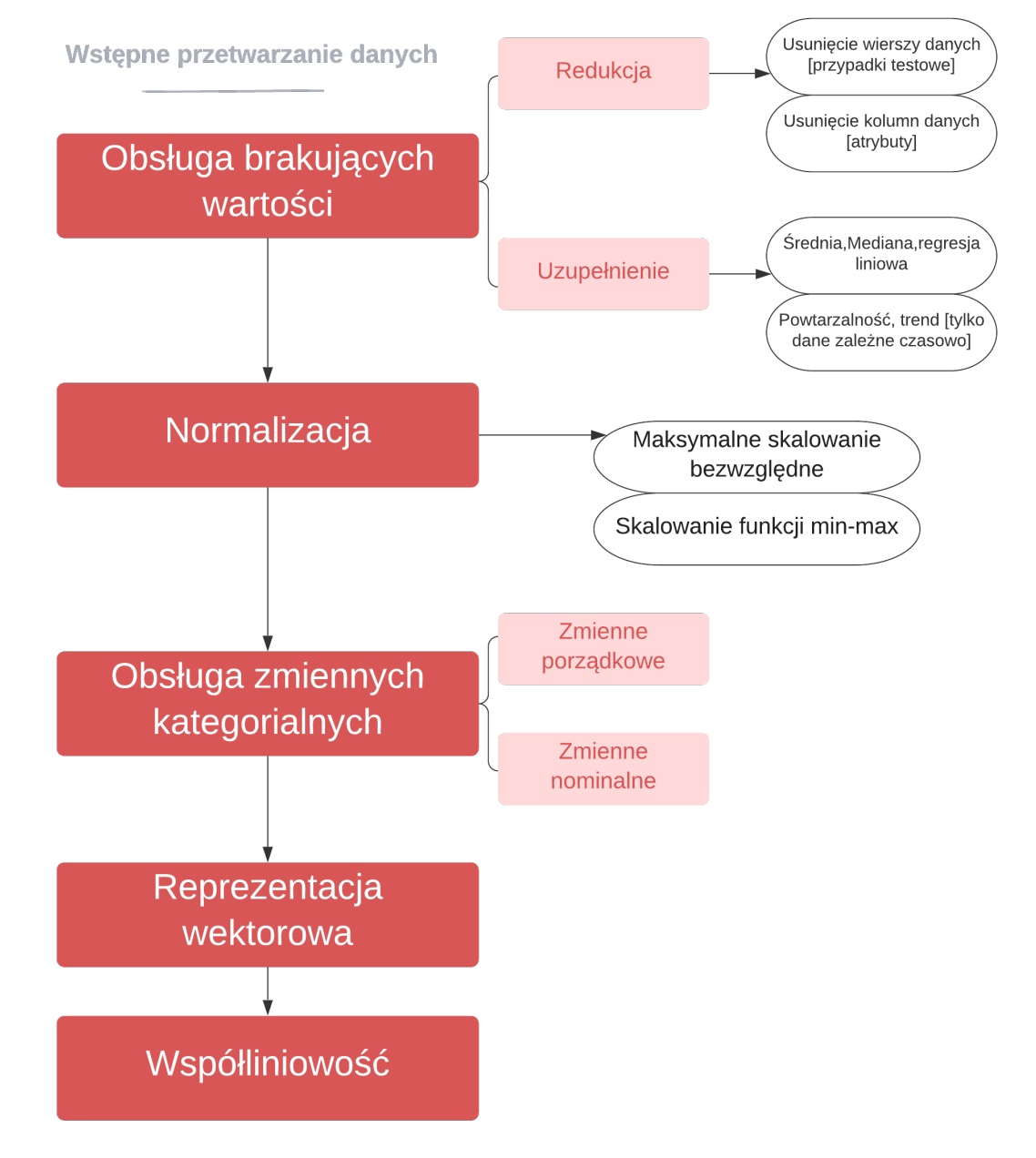

Figure 4: Schemat 1

Obsługa brakujących wartości

Możliwości obsługi brakujących wartości są jak już przedstawiono powyżej sa 2 : mniej polecana ze względu na utrate danych, redukcja dataset'u 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:

```
dataframe.dropna()
```

Parametryzując:

- axis — axis=0 jeśli chcemy usunąć wiersze lub axis=1 dla kolumn,
- how — dla how = 'all', wiersze i kolumny zostaną usunięte tylko wprzypadku, gdy wszystkie wartości kolumny lub wiersza to NaN. Domyślnie how jest ustawione na 'any' i skutkuje to usunięciem wiesza/kolumny z jakimi kolwiek pustymi wartościami.

Uzupełnienie danych inaczej imputacja, rozwiązuje problem w mniej stratny sposób i tak samo jak do redukcji są już gotowe rozwiązania w bibliotece sklearn.

```
imputed_mean = SimpleImputer(strategy="mean", missing_values=numpy.NaN,
 fill_value=-1)

imputed_median = SimpleImputer(strategy="median", missing_values=numpy.NaN,
 fill_value=-1)

imputed_most_frequent = SimpleImputer(strategy="most_frequent",
missing_values=numpy.NaN, fill_value=-1)

imputed_most_constant = SimpleImputer(strategy="constant",
missing_values=numpy.NaN, fill_value=-1)
```

Powyżej przedstawiono 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

Przekształcenie 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

[TODO] wstawić wzór podejmujemy średnią i dzielimy ją przez odchylenie standardowe.

[TODO] prezentacja wizualna Z praktycznego umieszczamy dane w zawężonym zakresie na osi.

```
def standarization(x_test, x_train):
    temp_x_train = x_train.loc[:, :].copy()
    temp_x_test = x_test.loc[:, :].copy()
    for iterator in ['age', 'trestbps', 'chol', 'restecg', 'thalach', 'oldpeak',
                     'slope', 'ca', 'thal']:
        scale = StandardScaler().fit(x_train[[iterator]])
        temp_x_train[iterator] = scale.transform(x_train[[iterator]])
        temp_x_test[iterator] = scale.transform(x_test[[iterator]])

    return temp_x_test, temp_x_train
```

Powyżej przedstawiono funkcję wykonującą standarycjację poprzez oblizenie średniej oraz odchylenia standardowego wykorzystując funkcję fit a następnie konwertując dane wykorzystując funkcję transform.

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 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.

[TODO] wizualizacja

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

[TODO] Wwykresy dla cech

Przy zastosowaniu reprezentacji wektorowej dla cech mocno od siebie uzależnionych zalecane jest zastosowanie :

```
drop_first=True
```

Zestawienie efektywności działania algorytmów

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 Numerical Python, zestaw narzędzi do obliczeń na macierzach, wektorach oraz o pakiet Science python umożliwiający metody numeryczne takie jak całkowanie, różniczkowanie itp. .
Do przygotowania danych wykorzystano zestaw narzędzi Pandas, ułatwiający tworzenie struktur danych i ich analizę. W celu wizualizacji wyników w postaci wykresów zastosowano Matplotlib. Część prezentacyjna czyli możliwość wprowadzenia danych w formularzu na stronie i weryfikacja wyniku dla wyuczonych już modeli wykorzystuje bibliotkę Flask.

Python

[^PYTHON]

[^PYTHON] https://www.python.org/downloads/release/python-390/

Flaks Templates

Numpy

Numpy to wydajny pakiet do obliczeń naukowych ,który idealnie nadaje się do pracy na tablicach na których wykonywane sa ciężkie operacje matematyczne.

SkitLearn

Scikit-learn to biblioteka implementująca algorytmy uczenia maszynowego. Sama biblioteka wykorzystuje :

- NumPy,
- Matplotlib,
- Pandas.

Matplotlib

JobLib

JobLib wykorzystany do zapisu wytrenowanych modeli poprzez swoje możliwości do operacji na obiektach Pythona.

Pandas

Pandas to pakiet ze struktura danych który udostępnia przydatne mechanizmy takie jak :

- zmiana rozmiarów na tabeli.

- segregacja zestawów danych.

Modułu projektu:

- *algorithms*:
  - *decisionForest* - implepmentacja algorytmu
  - *KNN* - implepmentacja algorytmu
  - *SVM* - implepmentacja algorytmu
- *data* - moduł odpowiada za wczytywanie i obróbkę danych testowych, oraz danych dostarczonych finalnie do weryfikacji modelu
- *doc* - praca oraz wszytkie dokumenty
- *result* - moduł odpowiedzialny za prezentację wyników w postaci wy kresów porównujących algorytmy oraz odpowiedzi na zadany problem

Trening algorytmu

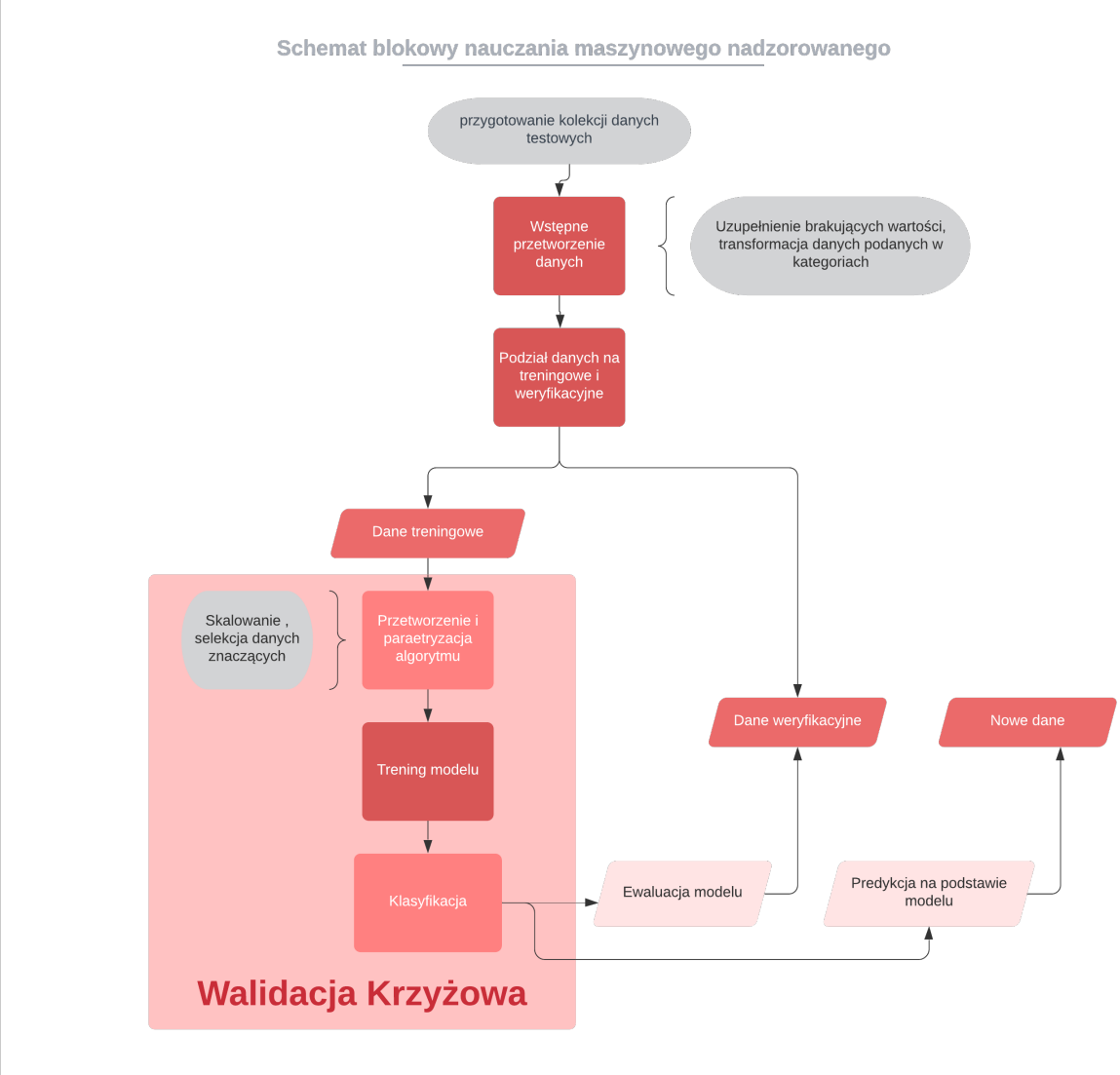

Figure 5: Schemat 1

Zgodnie z powyższym schematem po przeprocesowaniu wejściowego dataset'u dane należy podzielić

na dane treingowe oraz ewaluacyjne. Powszechnie stosowana K krzyżowa walidacja umożliwia maksymalne wykorzystanie dostarczonego wejśia do dostrajania paraetrów modelu.

K-krotna walidacja krzyżowa (ang. *Fold Cross-Validation*) to metodyka weryfikacji poprawności modeli nauczania masznowego. Opiera się ona na wyporze wartości swojego hiperparamtru jakim jest K, które może przyjąć dowolną wartość mniejszą lub równą od rozmiaru danych.

Po wyborze hiperparametru następuje segmentacjia danych na K jendakowej wielkości zestawów. Wykonywanych jest k iteracji, w każdej z nich na k-1 kolekcjach model jest trenowany, a na pozostałej jednej weryfikowany. Procedura efektywnie pomaga ocenić poprawność działania modelu i zastosowanego algorytmu.

[TODO] UZUPEŁNIENIE

Wybrane algorytmy uczenia maszynowego nadzorowanego

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 następuję rozróżnienie. W modelu kluczowa jest kolejnośc cech, które występują po sobie ponieważ determinuje to otrzymany rezultat.

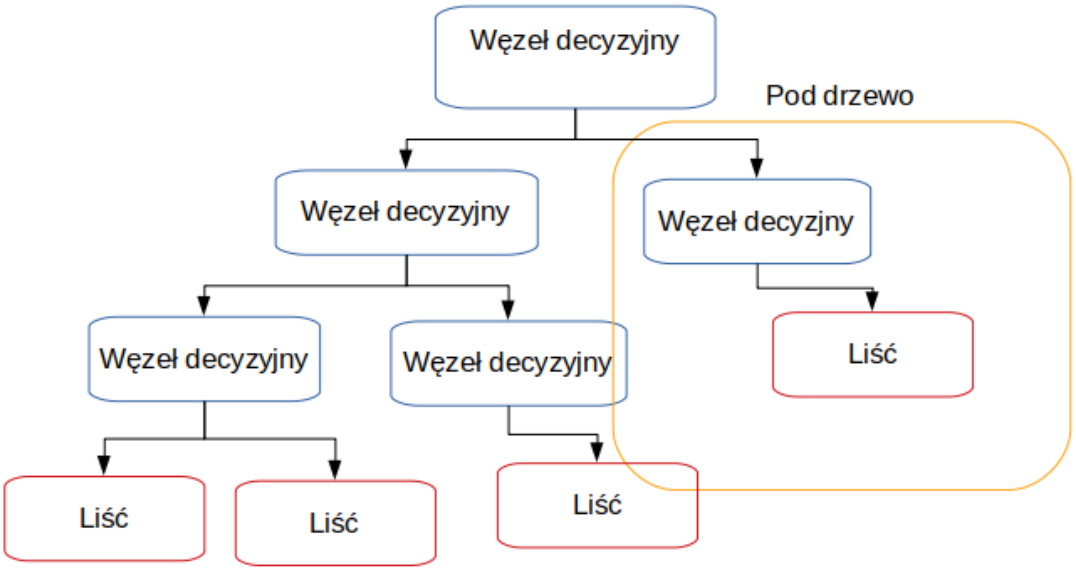

Figure 6: Schemat 1

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,którestaja się węzłami decyzyjnymi, dzielącymi rekordy na dwa mniejsze zestawy i tak aż nie ma możliwości dalszego podziału.

O 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.

W tym celu dla loswoych lasów decyzyjnych najpierw stosuję się **agregację bootstrap'ową**.

Z treningowego zestawu danch losuję się, co ważne 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.

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.

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

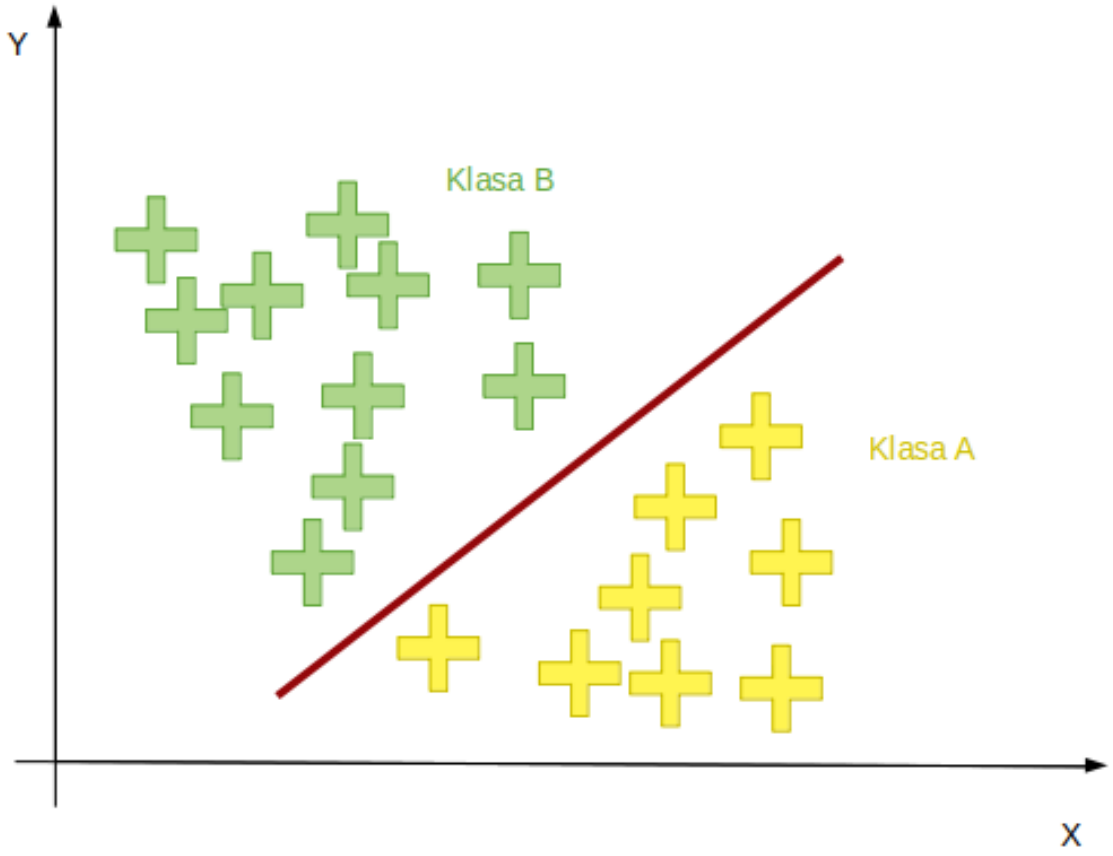

Figure 7: Schemat 2 [3]

Techinka ta gwarantuje precyzyjniejsze rezulatay niż drzewa deczyjne, 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.

Zbór dancyh wykorzytany w pracy nie jest aż tak kolosalny by zaszkodzić wydajności, a małym kosztem można uzyskać celnośc rozwiązania zadanego problemu: wykrwania występowania chorób

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

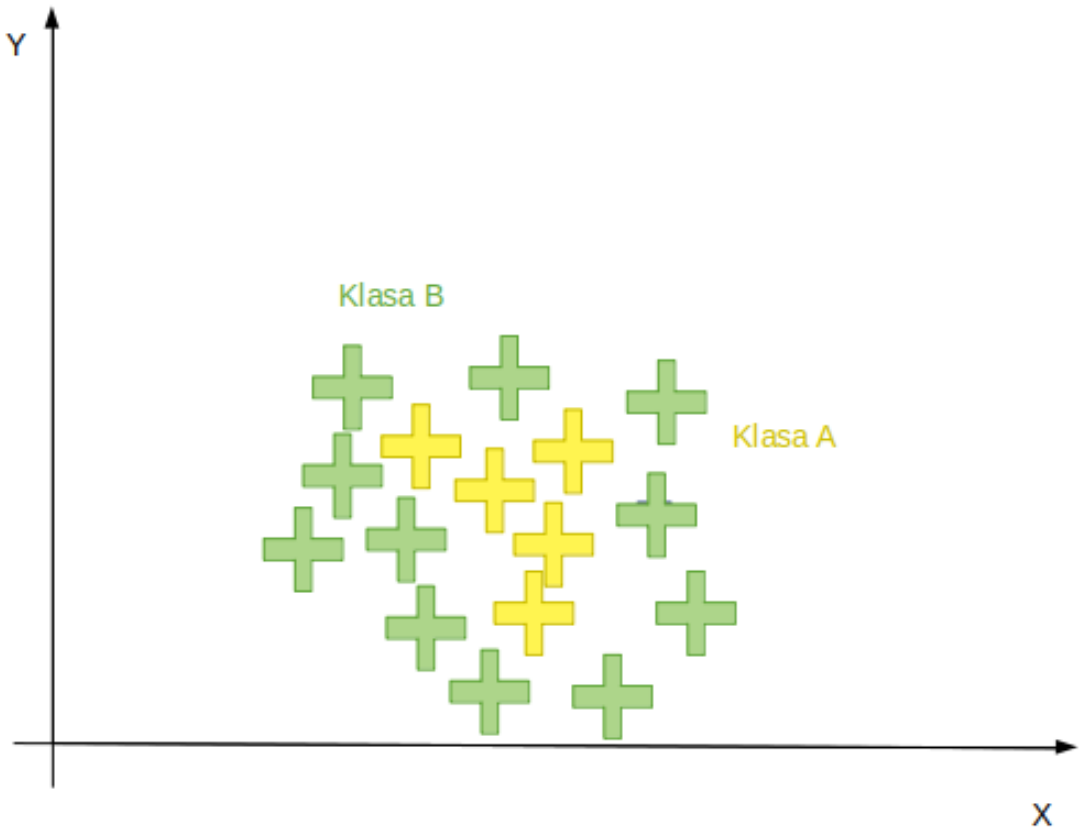

Figure 8: Schemat 2 [4]

sercowo-naczyniowych. Istnieje jednak ryzyko uzyskania rozkładu wartości który wyklucza graficzną frgmentacje zestawu danych na dwie części za pomocą prostej.

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, miasto(in. Manhattan) oraz Mińkowskiego.

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.

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.

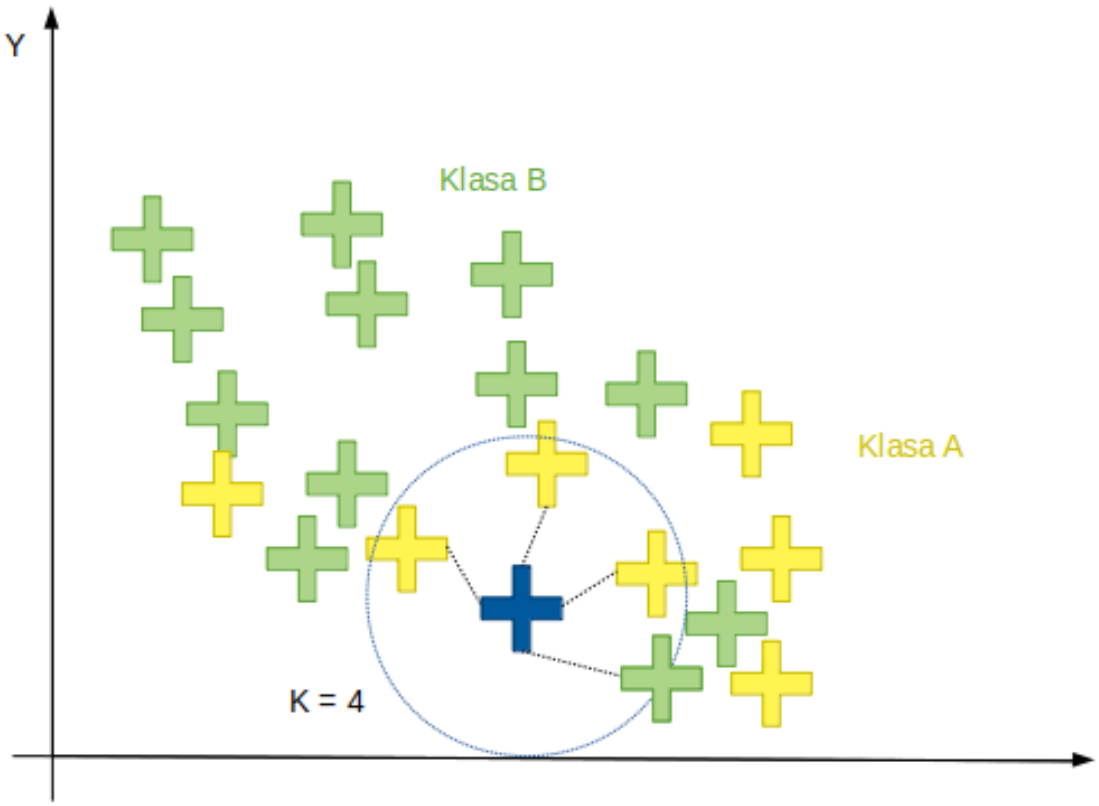

Figure 9: Schemat 3[5]

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.

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

!część niegotowa ze względu na braki implementacyjne !

Budowa modelu

Implementacja algorytmu 1: Losowe lasy decyzyjne

Implementacja algorytmu 2: Metoda wektorów nośnych

Implementacja algorytmu 3: K najbliższych sąsiadów

Wnioski i walidacja rozwiązania

Algorytm 1:Resultaty wnioski : Losowe lasy decyzyjne

Algorytm 2: Rezultaty wnioski: Metoda wektorów nośnych

Algorytm 3 : Rezultaty wnioski: K najbliższych sąsiadów

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

Podsumowanie i opisanie wpływu danych na model

[todo] porównanie do dnych statystycznych

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{https://ichi.pro/pl/uczenie-maszynowe-proste-wprowadzenie-96150019624312}
- @article{https://zpjn.wmi.amu.edu.pl/wp-content/uploads/2019/10/praca_magisterska.pdf t}
- @article{https://pdf.helion.pl/alguma/alguma.pdf}
- @article{https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d}
- @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{file:///C:/Users/User/Downloads/od_pojedynczych_drzew_do_losowego_lasu.pdf}
- @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/&httpsred
- @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%20l
- @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_ppc&utm_medium=cpc&utm_campaignid=1455363063&utm_adgroupid=650

392016246653:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=1011615&gclid=Cj0KCQiA0eC
- @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_ucz 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%
- @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/Cardiovascular-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{https://www.geeksforgeeks.org/ml-types-learning-supervised-learning/}
- @article{http://www.cs.ucr.edu/~mwile001/papers/thesis.pdf}
- @article{https://pl.wikipedia.org/wiki/Las_losowy}
- @article{https://python-graph-gallery.com/111-custom-correlogram/}
- @article{https://python-graph-gallery.com/242-area-chart-and-faceting/}
- @article{https://en.wikipedia.org/wiki/Random_forest}
- @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://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/}
- @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://en.wikipedia.org/wiki/Ejection_fraction}
- @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%
- @article{ Citation Request:

The authors of the databases have requested that any publications resulting from the use of the Data include the names of the principal investigator responsible for the Data collection at each institution. They would be: 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D. }