WSI - ćwiczenie 4

Agnieszka Wójtowicz

22 kwietnia 2022

Spis treści

1 Polecenie oraz elementy języka

1.1 Polecenie Prowadzącego

Tematem czwartych ćwiczeń jest regresja i klasyfikacja. W ramach tego zadania będą musieli Państwo zaimplementować algorytm ID3 i przeprowadzić klasyfikację dla zadanego zbioru danych.

1.2 Elementy języka

Wykorzystano język programowania Python w wersji 3.8. Skorzystano z pomocniczych bibliotek: *numpy 1.20.1*, *pandas 1.3.4*, *anytree 2.8.0*.

2 Cel zadania

Celem zadania jest implementacja algorytmu ID3 oraz wykorzystanie go do zbudowania drzewa na zadanym zbiorze danych.

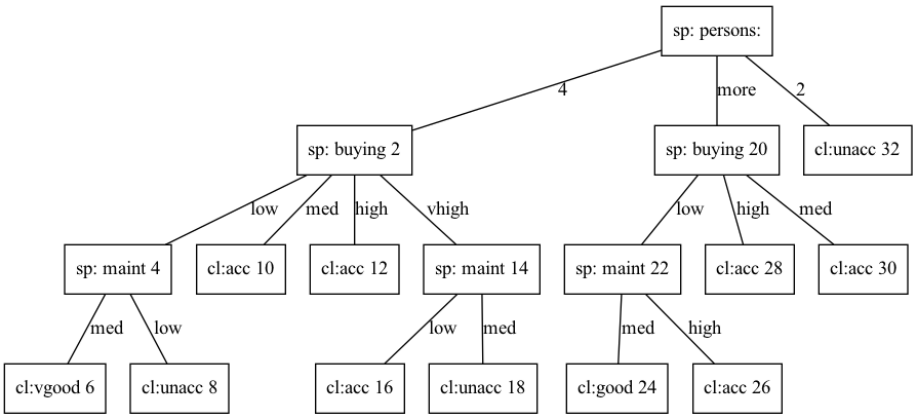
3 Wstęp

Zaimplementowano algorytm ID3 i wykorzystano go do klasyfikacji danych. Przyjęto klasyczną wersje algorytmu, dodatkowo implementując funkcjonalność polegającą na radzeniu sobie z tzw. "missing values". Przyjęto, że przykład z wartością cechy otrzymuje klasę dominującą wśród przykładów, które dotarły do węzła, w którym rozważana jest cecha o nieznanej drzewu wartości.

4 Wyniki

4.1 Przykładowe drzewo

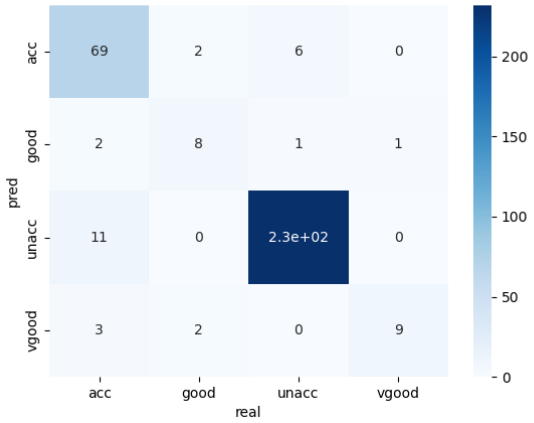
Rysunek 3 przedstawia przykładowe drzewo zbudowane dla bardzo małego podzbioru danych. Jest ono zdecydowanie mniej złożone niż końcowe drzewa uzyskiwane dla dalszych zbiorów treningowych, jednak nie zdecydowano się ich zamieszczać, ze względu na duże skomplikowanie.



Rysunek 1: Przykładowe drzewo ID3.

4.2 Przykładowy model wraz z jego oceną

Wykonano trening i test modelu dla podziału na zbiór treningowy i testowy w stosunku 8:2. Poniższa tabela oraz rysunek przedstawiają miary oceny jakości modelu oraz macierz pomyłek.



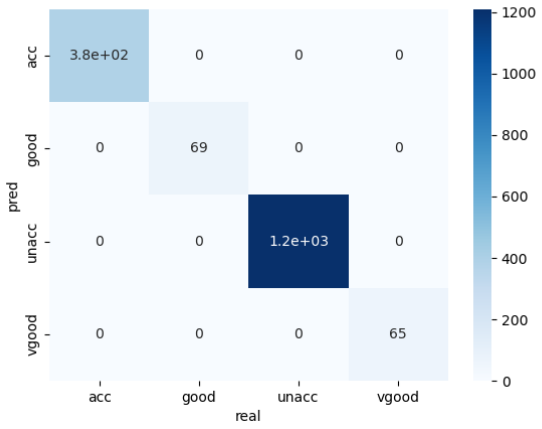
Rysunek 2: Macierz pomyłek modelu.

class	precision	recall	accuracy	f1-score
acc	0.89	0.81	0.93	0.85
good	0.8	0.8	0.98	0.8
unacc	0.95	0.97	0.95	0.96
vgood	0.8	0.92	0.99	0.86

Tabela 1: Miary oceny jakości modelu.

4.3 Model wytrenowany na całych danych

Wykonano trening i test modelu bez podziału zbioru (tzn trenowano model na całym dostępnym zbiorze). Poniższa tabela oraz rysunek przedstawiają miary oceny jakości modelu oraz macierz pomyłek.



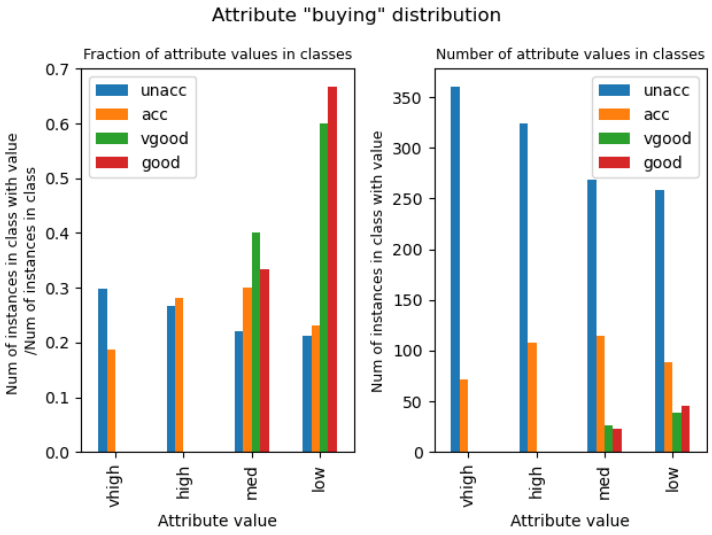
Rysunek 3: Macierz pomyłek modelu.

class	precision	recall	accuracy	f1-score
acc	1.0	1.0	1.0	1.0
good	1.0	1.0	1.0	1.0
unacc	1.0	1.0	1.0	1.0
vgood	1.0	1.0	1.0	1.0

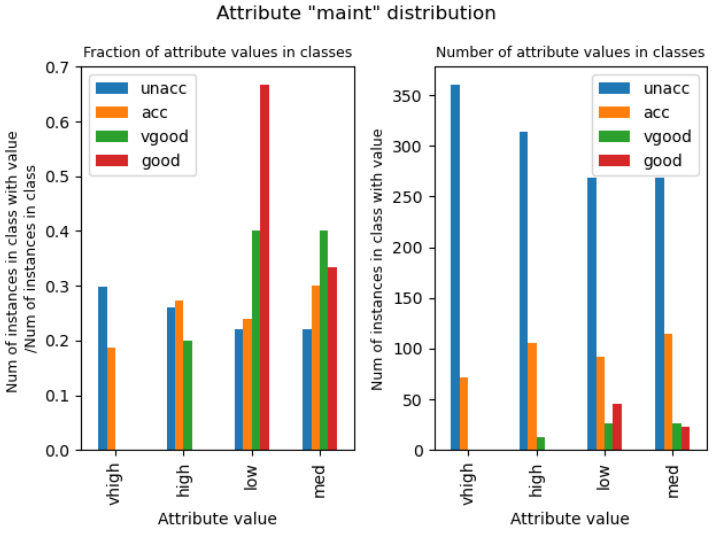
4.4 Analiza danych

Zbadano które wartości atrybutów występują najczęściej w danych klasach. Ze względu na dużą nierównowagę w liczności klas zbadano również jak duży jest procentowy udział przypadków posiadających daną wartość atrybutu w stosunku do liczności klasy. Może to dać jaśniejszą informacje na temat charakterystycznych dla danych klas wartości.

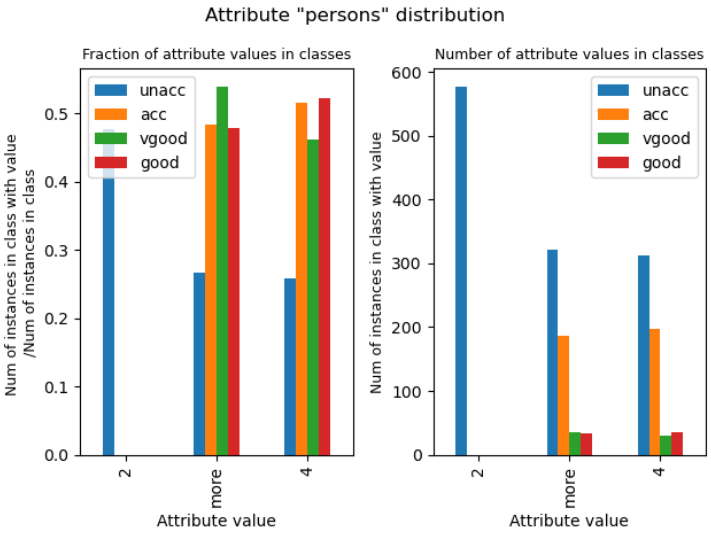
Histogramy wartości kolejnych atrybutów znajdują się na poniższych rysunkach.



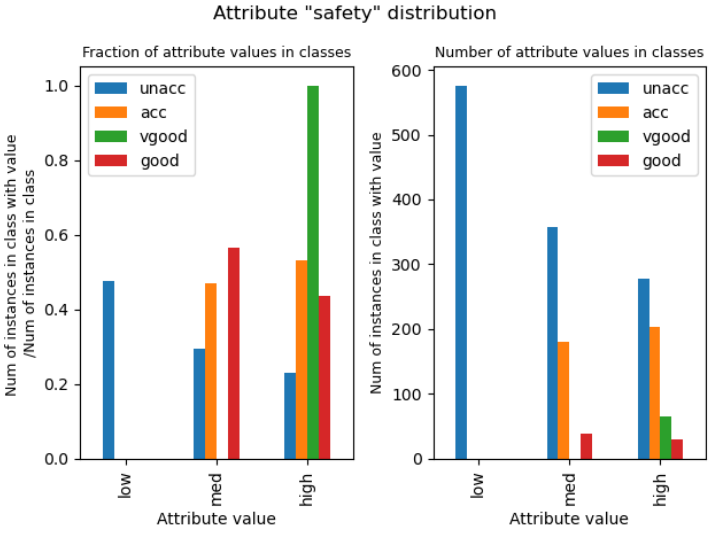
Rysunek 4: Rozkład wartości atrybutu ”buying”.



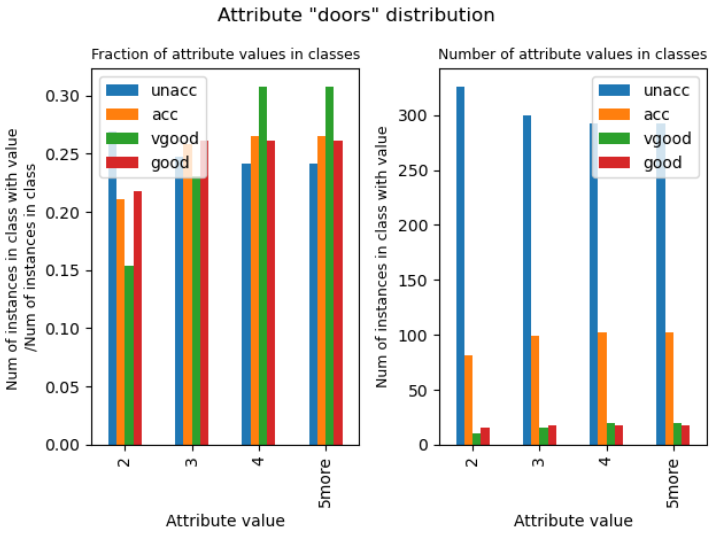
Rysunek 5: Rozkład wartości atrybutu ”maint”.



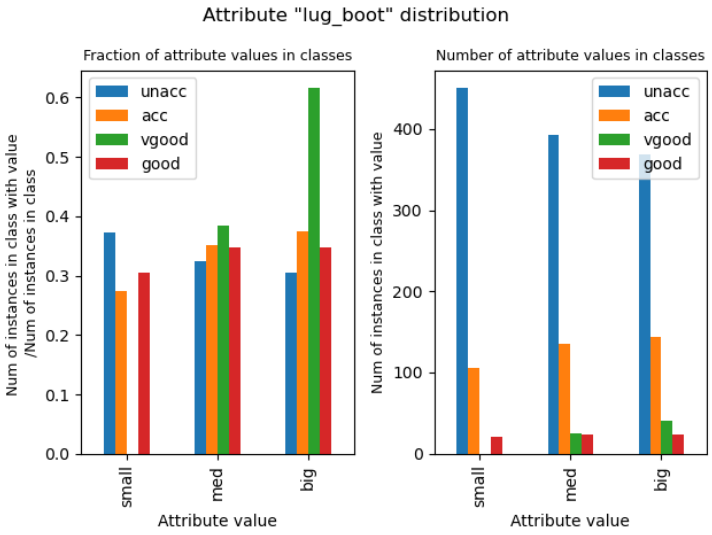
Rysunek 6: Rozkład wartości atrybutu ”persons”.



Rysunek 7: Rozkład wartości atrybutu ”safety”.



Rysunek 8: Rozkład wartości atrybutu ”doors”.



Rysunek 9: Rozkład wartości atrybutu ”lug_boot”.

4.5 Dane posortowane i nieposortowane

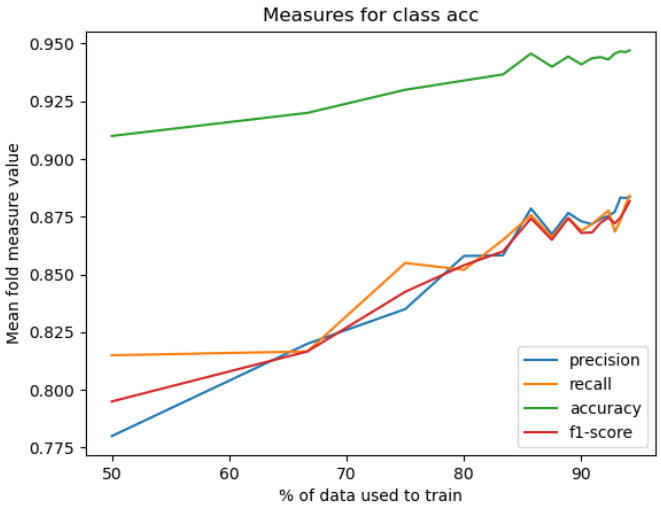
Drzewo decyzyjne jest algorytmem, w przypadku którego sortowanie nie ma wpływu na wynik końcowy. Budowanie kolejnych węzłów odbywa się na podstawie wyznaczania entropii danych, w którym to procesie kolejność danych nie jest istotna.

Moment w którym sortowanie danych mogłoby być istotne to gdyby zostały one posortowane przed podzieleniem na zbiór danych treningowy i testowy, a następnie dobór przykładów treningowych nie byłby losowy. Wtedy zbiór danych byłby mało reprezentatywny (np. pojawiłyby się przykłady tylko jednej klasy) i model miałby złe wyniki. Wykonanie takiej procedury wydaje się jednak bardzo specyficznym doborem danych, które na celu ma łatwy do przewidzenia rezultat - stworzenie złego modelu. Nie zdecydowano się więc przeprowadzić takiej procedury.

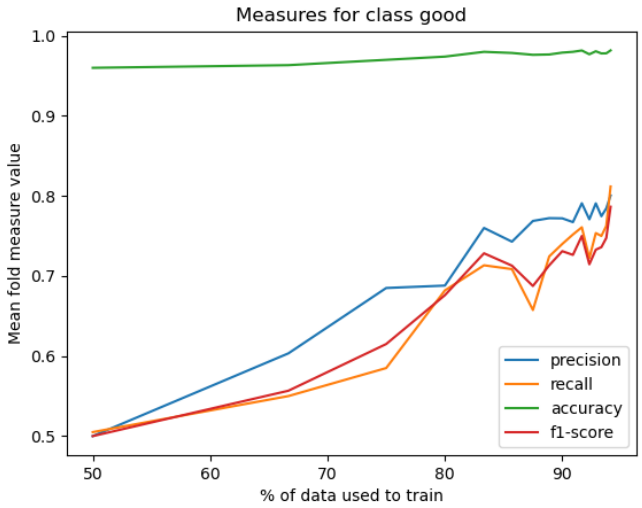
4.6 Podział na zbiór treningowy i testowy

Model trenowano na zbiorze wykorzystując procedurę walidacji krzyżowej. Zmienianie jej krotności powoduje zmianę proporcji zbioru treningowego i testowego w procesie trenowania i testowania modelu.

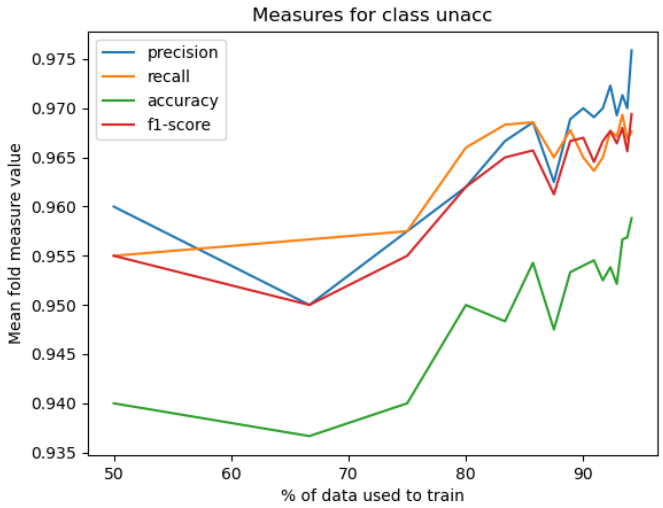
Poniższe rysunki ilustrują zmiany wartości zaimplementowanych ocen jakości w zależności od procentowego udziału zbioru treningowego. Oddzielnie przedstawiono miary jakości dla 4 klas.



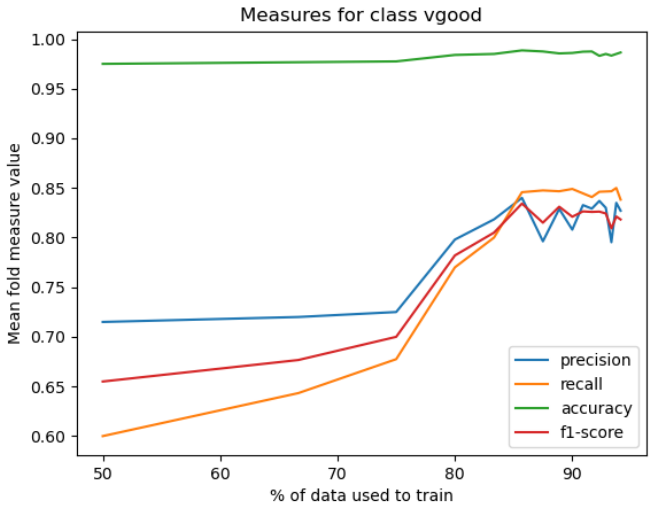
Rysunek 10: Miary oceny jakości w zależności od wielkości zbioru treningowego - klasa acc.



Rysunek 11: Miary oceny jakości w zależności od wielkości zbioru treningowego - klasa good.



Rysunek 12: Miary oceny jakości w zależności od wielkości zbioru treningowego - klasa unacc.



Rysunek 13: Miary oceny jakości w zależności od wielkości zbioru treningowego - klasa vgood.

5 Wnioski

5.1 Analiza danych

Wstępna analiza danych wykazała, że pewne wartości atrubutów są charakterysyczne dla danych klas. Wśród nich wskazać warto:

1. cecha "persons", wartość "2", charakterystyczna dla klasy "unacc" (6)
2. cecha "safety", wartość "low", charakterystyczna dla klasy "unacc" (7),
3. cecha "safety", wartość "high", mają ją wszystkie przykłady klasy "vgood" (7).

Podobnych zależności można szukać dalej. Na rysunku 13 widać, że mimo iż drzewo wytrenowane zostało na bardzo niewielkim procencie danych, wspomniana zależność (persons == 2 - klasa "unacc") została została dostrzeżona przez drzewo podczas treningu.

5.2 Podział na zbiór treningowy i testowy

Model wytrenowany na całym zbiorze treningowym osiągnął na nim idealne miary jakości. Analizowany zbiór danych jest więc idealnie, jednoznacznie separowalny na klasy na podstawie danych. Dzieląc zbiór na treningowy i testowy nie będziemy się więc spodziewać zjawiska przeuczenia, a raczej niedouczenia modelu. Wszystkie rysunki ukazujące wartości miar jakości modelu w zależności od krotności walidacji jasno wykazują (dla wszystkich klas) ich poprawę wraz ze wzrostem procentowego udziału zbioru treningowego w całym zbiorze. Mając do dyspozycji niewielką próbkę danych model klasyfikuje nowo przybyłe

dane testowe słabo (szczególnie w przypadku mniej licznych klas). Można więc mówić tu o niedouczeniu, a model wykazuje słabą generalizacje. Dostając większą (a więc zwykle i bardziej reprezentatywną) próbkę, klasyfikuje dane lepiej, choć wzrost miar ocen jakości jest od pewnego poziomu jest mniej widoczny (ponad wartość 90 procent).