

Analiza danych sieciowych

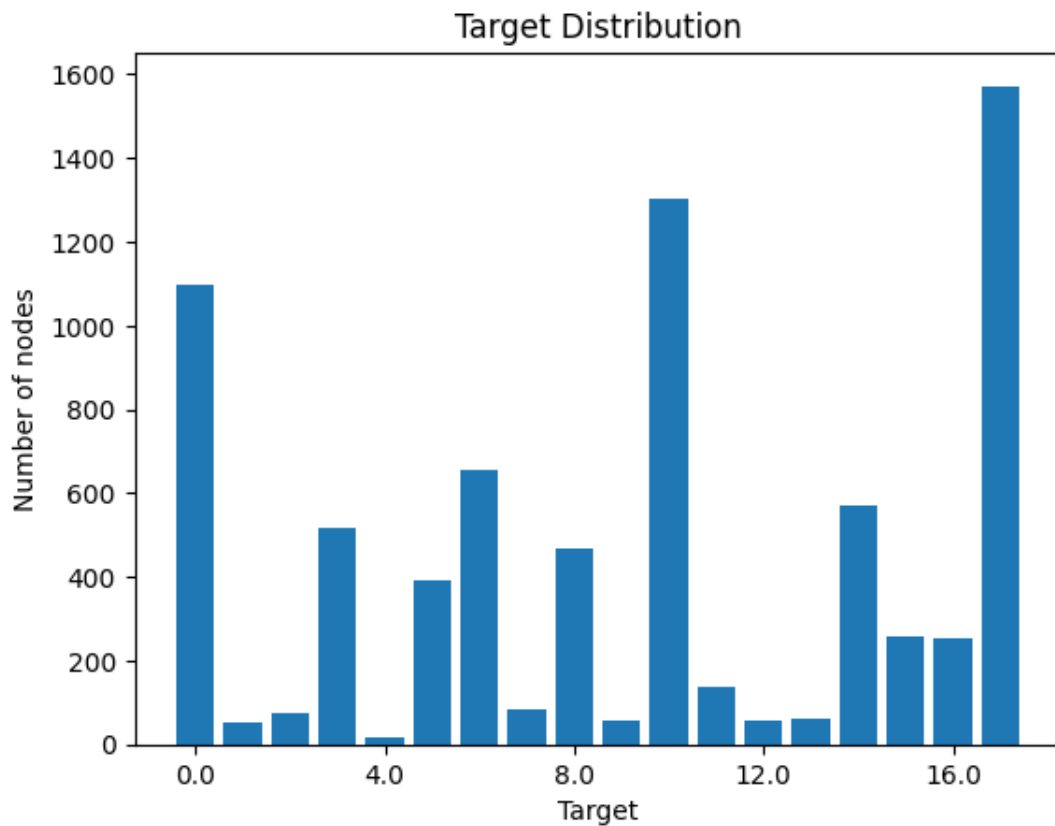
Jakub Gałka 250138

Maj 2023

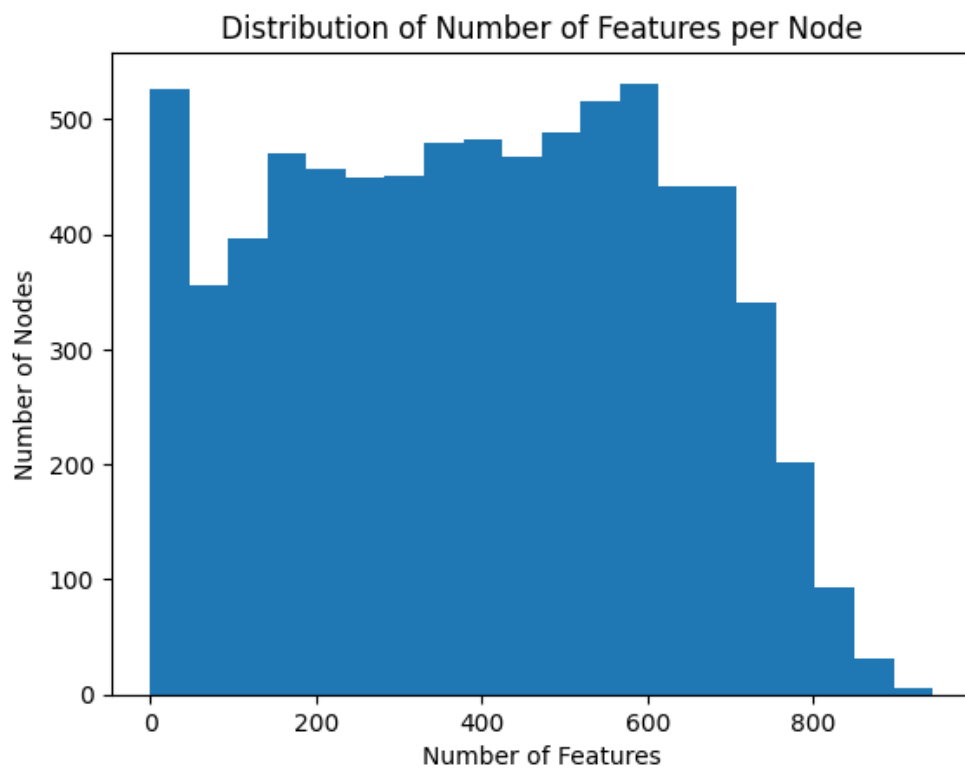
Zbiór danych

1. Węzły to użytkownicy LastFM z krajów azjatyckich.
2. Krawędzie to wzajemne relacje obserwowania między użytkownikami.
3. Każdy z węzłów ma przypisana etykiety oznaczające kraj użytkownika.
4. Cechy wierzchołków są wydobywane na podstawie artystów lubianych przez użytkowników.
5. Klasyfikacja węzłów - trzeba przewidzieć lokalizację użytkowników.

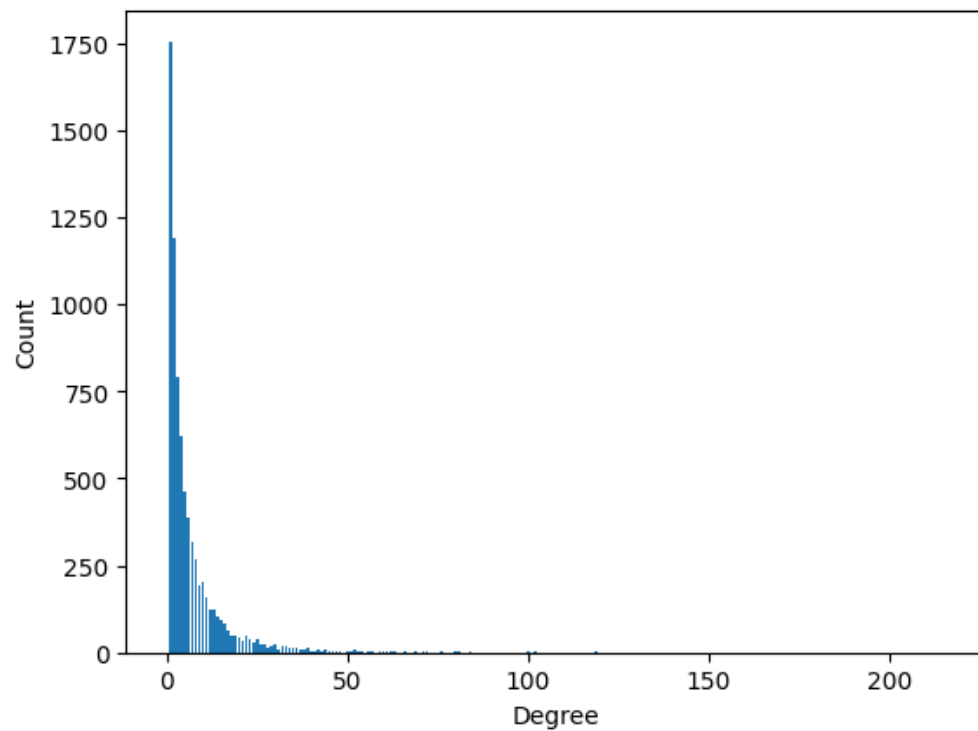
Etykety



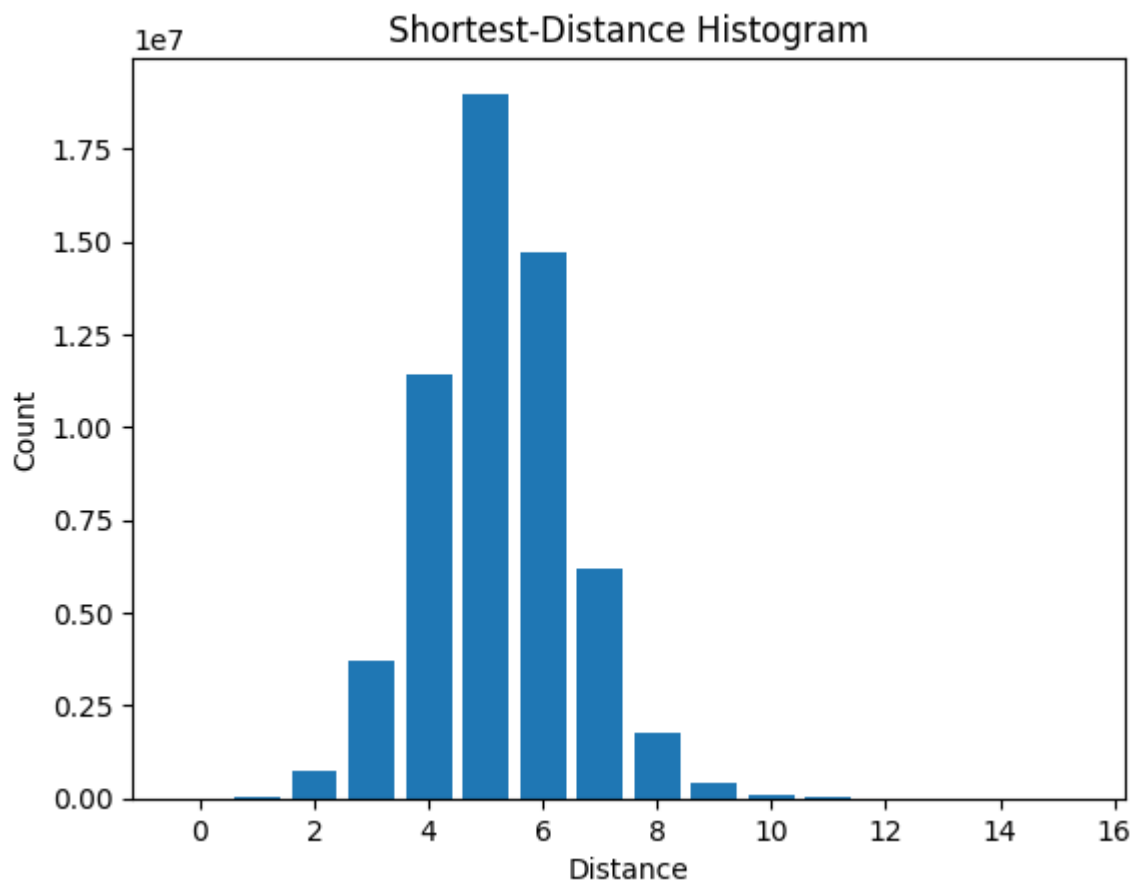
Cechy



Stopnie wierzchołkow

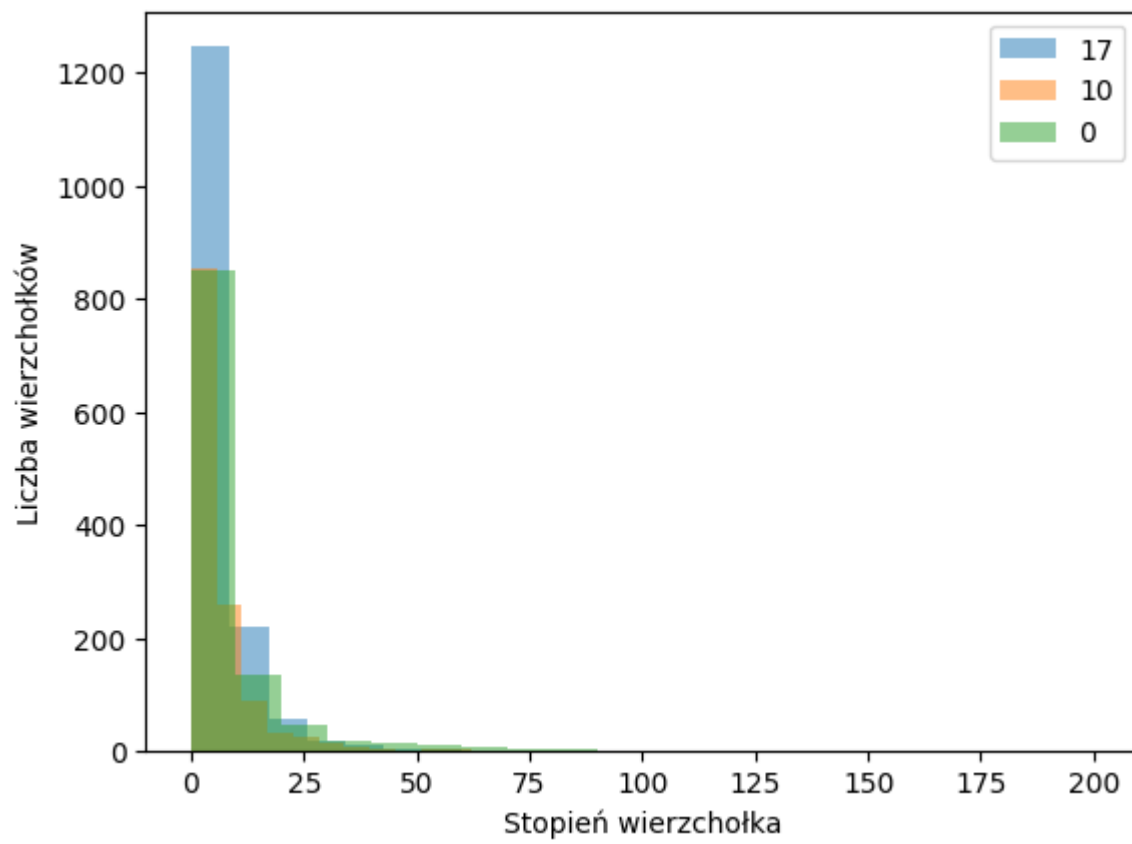


Najkrótsze ścieżki



Statystyki opisowe dla podgrafu 3 najliczniejszych klas

- Liczba węzłów: 3843
- Liczba krawędzi: 13835
- Średni stopień wierzchołka: 7.200104085349987
- Liczba składowych spójnych: 18



Średni współczynnik klastrowania dla poszczególnych klas:

- 0: 0.240550
- 10: 0.233860
- 17: 0.168278

Random forrest

Class	Precision	Recall	F1-Score	Support
0	0.93	0.81	0.87	124
1	0.00	0.00	0.00	5
2	0.00	0.00	0.00	7
3	0.71	0.88	0.79	52
4	0.00	0.00	0.00	2
5	0.84	0.61	0.71	44
6	0.54	0.46	0.50	69
7	0.00	0.00	0.00	7
8	0.93	0.82	0.87	50
9	0.00	0.00	0.00	4
10	0.66	0.97	0.79	130
11	1.00	0.33	0.50	12
12	0.00	0.00	0.00	4
13	1.00	0.17	0.29	6
14	0.77	0.53	0.63	51
15	1.00	0.61	0.76	31
16	1.00	0.17	0.30	23
17	0.64	0.91	0.75	142
Accuracy			0.73	763
Macro Avg	0.56	0.41	0.43	763
Weighted Avg	0.74	0.73	0.70	763

Centroids

- Dla każdej klasy oblicz centroid, który jest średnią wartością cech dla wszystkich elementów należących do danej klasy.
- Dla każdego elementu ze zbioru testowego - oblicz odległość kosinusowa wektora cech tego elementu od każdego z centroidów dla wszystkich klas.
- Przypisz elementowi klasę odpowiadającą najbliższemu centroidowi.

Class	Precision	Recall	F1-Score	Support
0	0.93	0.56	0.70	124
1	0.05	0.20	0.08	5
2	0.06	0.57	0.10	7
3	0.69	0.38	0.49	52
4	0.00	0.00	0.00	2
5	0.62	0.45	0.53	44
6	0.42	0.39	0.41	69
7	0.08	0.14	0.11	7
8	0.90	0.56	0.69	50
9	0.04	0.25	0.06	4
10	0.74	0.42	0.53	130
11	0.35	0.58	0.44	12
12	0.00	0.00	0.00	4
13	0.06	0.50	0.10	6
14	0.66	0.37	0.48	51
15	0.55	0.68	0.61	31
16	0.15	0.43	0.22	23
17	0.67	0.48	0.56	142
Accuracy			0.46	763
Macro Avg	0.39	0.39	0.34	763
Weighted Avg	0.66	0.46	0.53	763

Analiza wyników

Precyzja: Reprezentuje zdolność klasyfikatora do poprawnego identyfikowania pozytywnych instancji dla danej klasy. W raporcie precyzja jest wyświetlana dla każdej klasy. Na przykład w klasie 0 precyzja wynosi 0,93, co wskazuje, że gdy klasyfikator przewiduje próbkę jako klasę 0, jest poprawny w 93% przypadków. Jednak dla klas 1 i 2 wynosi 0,00, co sugeruje, że klasyfikator nie przewiduje poprawnie tych klas.

Recall: Mierzy zdolność klasyfikatora do poprawnego identyfikowania pozytywnych instancji spośród wszystkich rzeczywistych pozytywnych instancji dla danej klasy. W raporcie recall jest pokazany dla każdej klasy. Na przykład, w klasie 0, odzysk wynosi 0,81, co wskazuje, że klasyfikator może zidentyfikować 81% rzeczywistych próbek klasy 0. Jednak w przypadku klas 1 i 2 wartość ta wynosi 0,00, co oznacza, że klasyfikator nie jest w stanie poprawnie zidentyfikować tych klas.

Wynik F1: Jest to średnia harmoniczna precyzji i wycofania i zapewnia równowagę między tymi dwoma wskaźnikami. Wynik F1 uwzględnia zarówno precyzję, jak i wycofanie, a wyższa wartość wskazuje na lepszą równowagę. W raporcie wynik F1 jest wyświetlany dla każdej klasy. Na przykład w klasie 0 wynik F1 wynosi 0,87, co stanowi rozsądną równowagę między precyzją a wycofaniem. Jednak dla klas 1 i 2 wynik F1 wynosi 0,00, co wskazuje na słabą wydajność dla tych klas.

Wsparcie: Oznacza liczbę próbek w każdej klasie w testowym zbiorze danych. W raporcie wsparcie jest wyświetlane dla każdej klasy. Na przykład klasa 0 ma wsparcie 124, co oznacza, że w testowym zbiorze danych znajduje się 124 próbek klasy 0. Podobnie, klasy 1 i 2 mają wsparcie tylko 5 i 7.

Wnioski

Model osiągnął wysoka precyzję dla klasy 0, 8, 10 i 17, co wskazuje na niski wskaźnik fałszywych pozytywów. Czulość jest wysoka dla klasy 0, 3, 10 i 17, co oznacza niski wskaźnik fałszywych negatywów. Miary F1 są stosunkowo wysokie dla klas 0, 3, 8, 10 i 17, sugerując dobrą równowagę między precyzją a czulością dla tych klas. Niektóre klasy (np. 1, 2, 4, 7, 9, 11, 12, 13 i 16) mają niską precyzję, czulość i miary F1, co wskazuje na słabą skuteczność modelu w przewidywaniu tych klas. Ogólna dokładność modelu na zbiorze danych wynosi 73%. Warto jednak zauważyć, że sama dokładność może nie dawać pełnego obrazu skuteczności modelu, zwłaszcza gdy zbiór danych jest nie zrównoważony.

W drugiej metodzie klasyfikacji: Precyzja, czulość i miary F1 są ogólnie niższe w porównaniu do pierwszej metody. Większość klas ma niską precyzję, czulość i miary F1, co wskazuje na słabą skuteczność modelu we wszystkich klasach. Dokładność modelu na tym zbiorze danych wynosi 46%. Jednakże, można zauważyć, że recall osiąga średnio lepsze wyniki dla większości klas w porównaniu do metody 1.