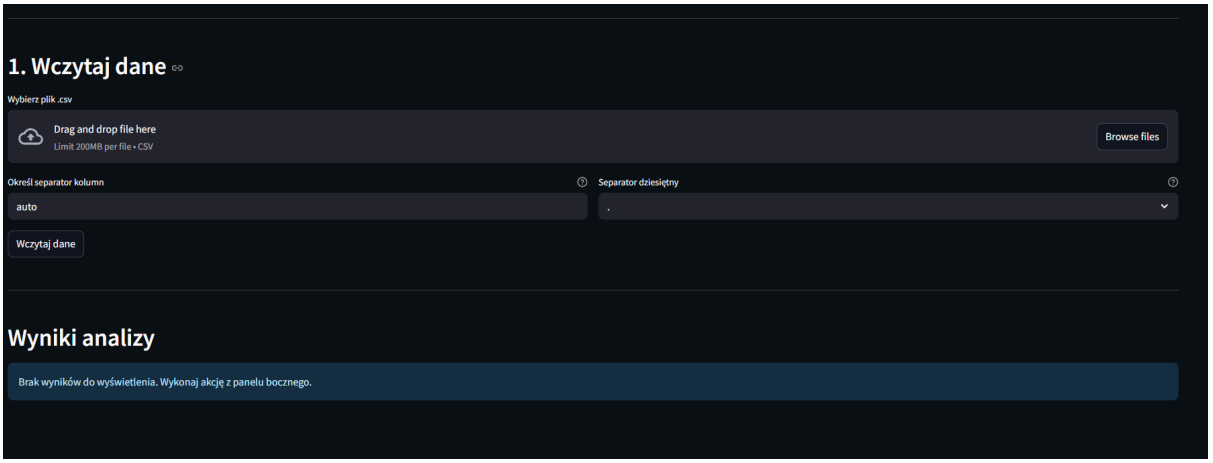
Data Visualization Tool with Streamlit

Radosław Mocarski

Tomasz Zapart

Projekt ten polegał na utworzeniu aplikacji webowej umożliwiającej przeprowadzenie wizualizacji danych. Użytkownikowi umożliwiono tworzenie wykresów m.in histogram, wykres pudełkowy, liniowy. Również możemy redukować wymiarowość naszych danych przy pomocy t-SNE, UMAP, TriMAP, PaCMAP. Dodatkowo istnieje możliwość policzenia korelacji (Pearsona, Spearmana), a także przeanalizowania podstawowych statystyk opisowych.

Wczytanie Danych

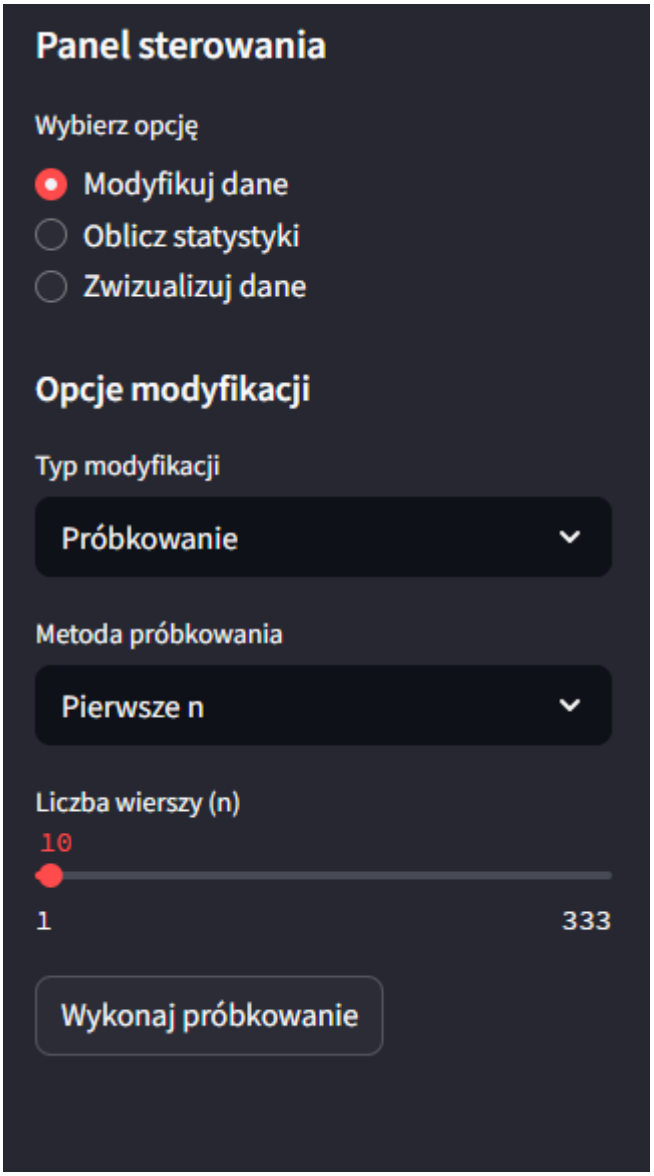


Aplikacja umożliwia wczytanie danych typu csv. Można określić separator kolumnowy (domyślnie automatycznie) oraz separator dziesiętny (domyślnie “.”). Możemy przeglądać nasz katalog w celu wybrania danych. Gdy już dokonamy wyboru należy nacisnąć przycisk wczytaj dane.
Po wczytaniu danych powinniśmy widzieć następująca strone, na której widzimy obecny podgląd danych.

Panel sterowania

Panel sterowania jest zorganizowany w trzy sekcje modyfikacja danych, analiza statystyczna, wizualizacja danych.
W sekcji modyfikacja danych możemy usuwać kolumny, w celu pozbycia się pewnych zmiennych, próbkowanie danych, umożliwiające działanie na mniejszych danych, oraz opcje redukcji wymiarowości z algorytmami (t-sne, Umap, PacMap, TriMap).
Sekcja analiza statystyczna zawiera opcje statystyki opisowe, która generuje tabelę z podstawowymi statystykami (liczebność, średnia, odchylenie standardowe, min, max, kwartyle) dla wszystkich kolumn numerycznych. Dodatkowo istnieje opcja obliczenia korelacji Pearsona oraz Spearmana.

Sekcja “wizualizacja danych” pozwala na generowanie różnorodnych wykresów takich jak wykres punktowy, liniowy, pudełkowy, mapa ciepła



Zbiór danych Palmer Penguins

Jest to zbiór danych opisujących podstawowe cechy pingwinów, jakie opisał badacz dr Kristen Gorman. Obserwacje prowadzono na trzech wyspach Archipelagu Palmera: Torgersen, Biscoe i Dream. Zbiór zawiera dane dotyczące 344 pingwinów, zebrane w latach 2007-2009.
Poniżej znajduje się lista:

- **species:** gatunek pingwina (Adelie, Chinstrap lub Gentoo).
- **island:** nazwa wyspy, na której zaobserwowano osobnika (Torgersen, Biscoe lub Dream).
- **bill_length_mm:** długość dzioba (jego górnej krawędzi, tzw. *culmen*) w milimetrach.
- **bill_depth_mm:** głębokość dzioba w milimetrach, mierzona w najgrubszym miejscu.
- **flipper_length_mm:** Długość płetwy (skrzydła) pingwina w milimetrach.

- **body_mass_g:** Masa ciała osobnika w gramach.
- **sex:** Płeć pingwina

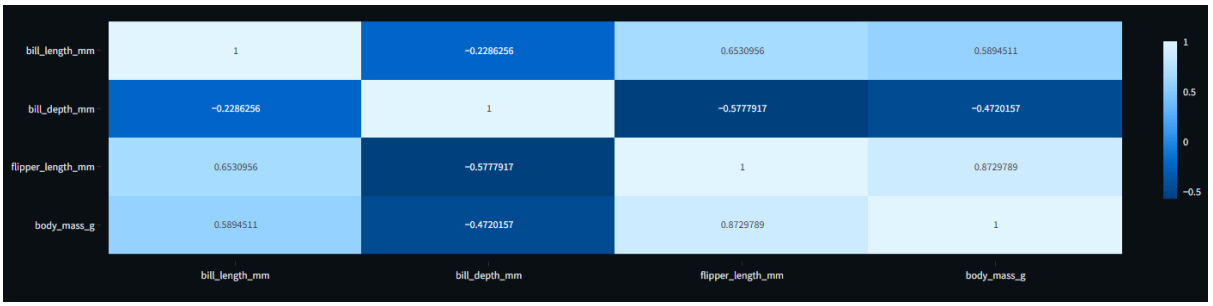
Statystyki opisowe

	count	mean	std	min	25%	50%	75%	max
bill_length_mm	333	43.9928	5.4687	32.1	39.5	44.5	48.6	59.6
bill_depth_mm	333	17.1649	1.9692	13.1	15.6	17.3	18.7	21.5
flipper_length_mm	333	200.967	14.0158	172	190	197	213	231
body_mass_g	333	4207.0571	805.2158	2700	3550	4050	4775	6300

Największe różnice widzimy w masie pingwinów, najmniejsze różnice w zauważamy w głębokości dzioba. Rozkład cech zdaje się być dość symetryczny, świadczą o tym podobne wartości mediany i średniej.

Korelacja liniowa pearsona

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
bill_length_mm	1	-0.2286	0.6531	0.5895
bill_depth_mm	-0.2286	1	-0.5778	-0.472
flipper_length_mm	0.6531	-0.5778	1	0.873
body_mass_g	0.5895	-0.472	0.873	1



Długość płetw, masa ciała i długość dzioba są ze sobą mocno powiązane. Jak jedna z nich rośnie, pozostałe cechy też rosną. Cechy te świadczą, o tym że większy pingwin będzie ma większe wymiary pozostałych cech, co jest normalne wśród zwierząt.
Większe pingwiny (cięższe, z dłuższymi płetwami) mają tendencję do posiadania bardziej smukłych dziobów.

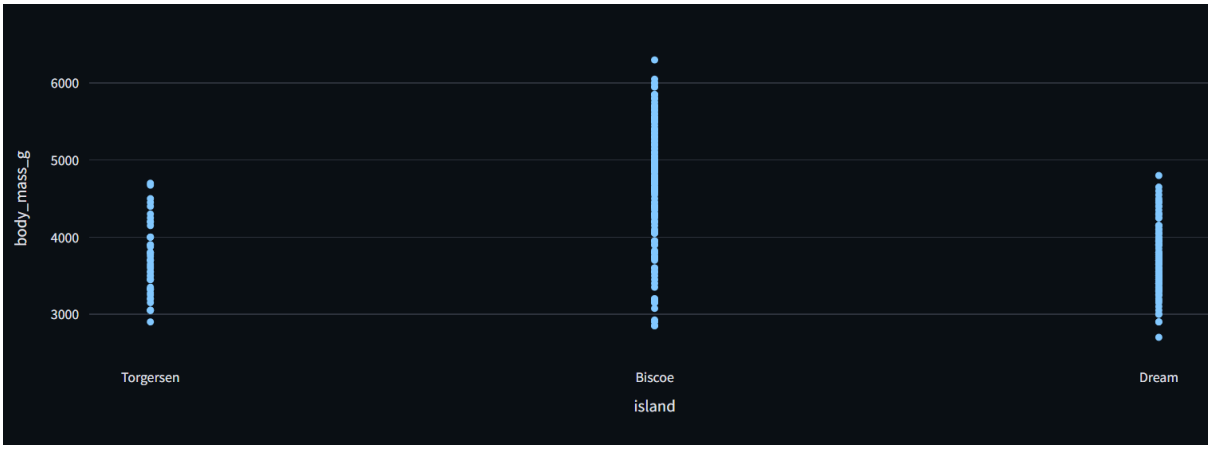
Gatunek pingwina, a masa

Widzimy, że gatunki pingwinów różnią się jeśli chodzi o wagę. W szczególności widzimy jak gatunek Gentoo istotnie więcej waży niż pozostałe gatunki.

Płeć, a masa

Widzimy, że samce więcej ważą niż samice, co jest też normalnym zjawiskiem.

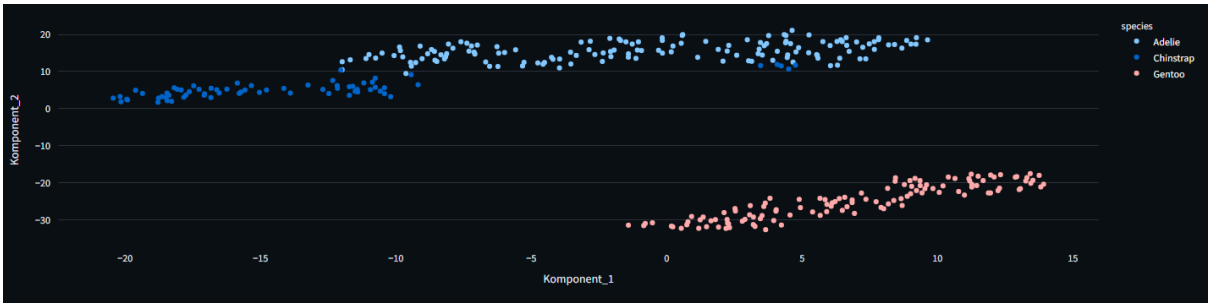
Badane wyspy, a masa pingwina



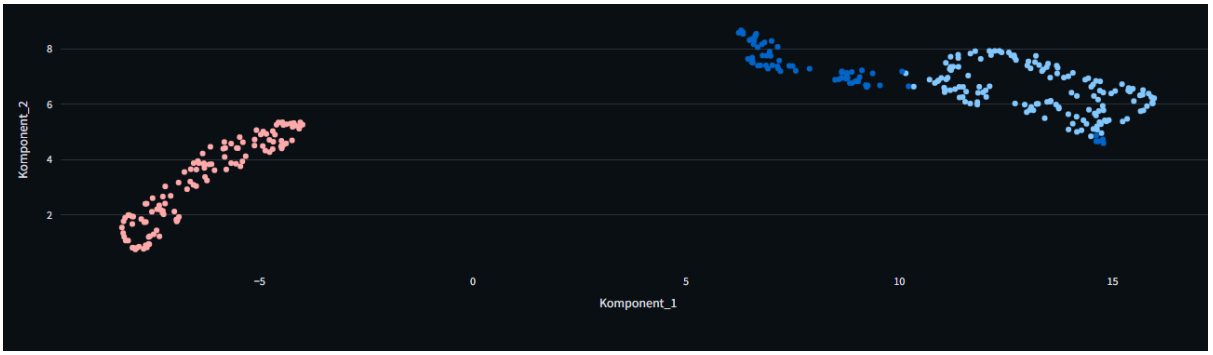
Widzimy, że wyspa Biscoe ma najbardziej rozproszone dane. Prawdopodobnie wynika to z faktu, iż występują tu zapewne dwa gatunki lub więcej, czyli Gentoo i Chinstrap.

Redukcje wymiarowości

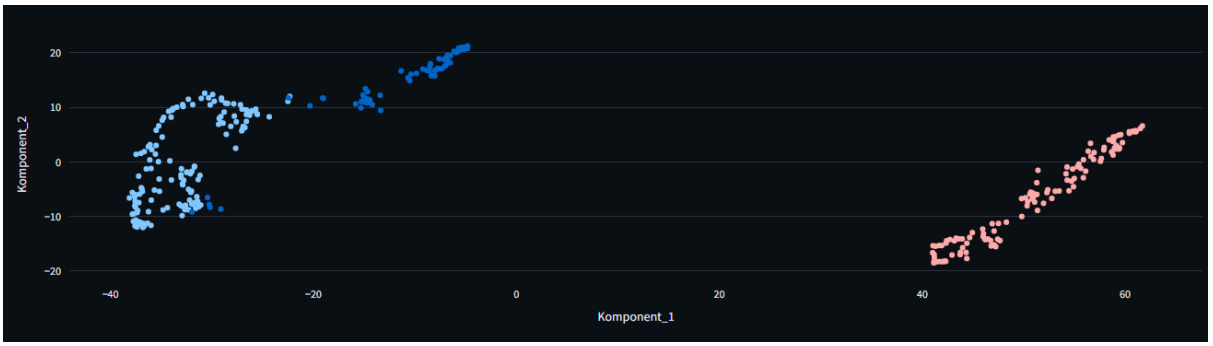
t-sne



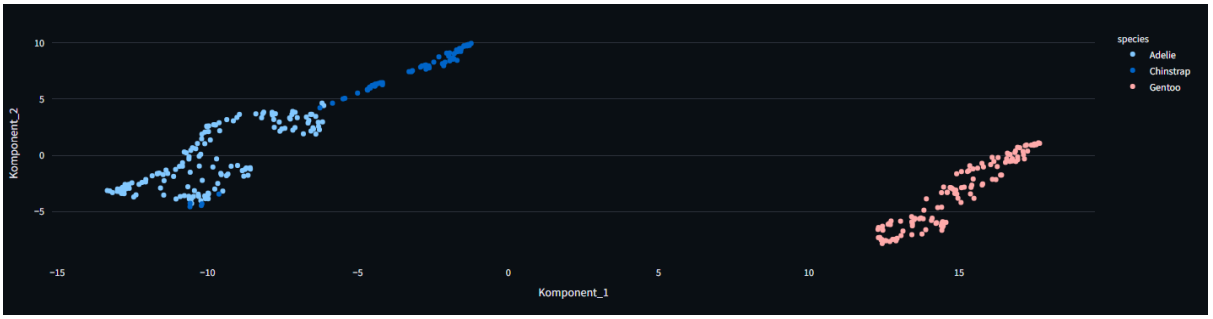
UMAP



TriMAP



PacMAP



Wszystkie algorytmy redukcji wymiarowości odseparowały gatunki. Jak widać najbardziej oddzielony został gatunek Gentoo, gdyż najbardziej różnił on się masą. Cecha ta była wysoce skorelowana z innymi parametrami, stąd też wszystkie algorytmy postanowiły w tak mocny sposób oddzielić te gatunki.

Zbiór danych Wine

Jest to zbiór zawierający wyniki analizy chemicznej win pochodzących od trzech różnych producentów we Włoszech. Zawiera 13 unikalnych cech opisujących właściwości chemiczne/fizyczne win.
Poniżej znajdują się zmienne opisujące wina:

- alcohol - zawartość alkoholu w winie, wyrażona procentowo.
- malic_acid - stężenie kwasu jabłkowego, jednego z głównych kwasów organicznych w winie. Wpływa na smak i kwasowość.
- ash - zawartość popiołu, który jest miarą całkowitej zawartości minerałów w winie. Otrzymuje się go przez odparowanie wina i spalenie pozostałości.
- alcalinity_of_ash - miara zasadowości popiołu, która również odzwierciedla zawartość minerałów, w szczególności metali alkalicznych
- magnesium - stężenie magnezu w winie
- total_phenols - całkowita ilość związków fenolowych. Fenole mają duży wpływ na cechy wina, takie jak kolor, smak, goryczka i potencjał starzenia
- flawonoidy - stężenie flawonoidów, które są specyficzną i bardzo ważną grupą fenoli. Odpowiadają za wiele właściwości antyoksydacyjnych, kolor i smak wina
- proanthocyanins - stężenie proantocyjanidyn, które są rodzajem skondensowanych tanin i wpływają na cierpkość oraz strukturę wina.
- color_intentisy - miara nasycenia barwy wina
- hue - odcień koloru wina
- od280/od315_of_diluted_wines - jest to stosunek absorbancji (pochłaniania światła) przy długościach fali 280 nm i 315 nm. Używany jako wskaźnik stężenia białek w stosunku do innych związków.
- proline - stężenie proliny

Statystyki opisowe

	count	mean	std	min	25%	50%	75%	max
alcohol	178	13.0006	0.8118	11.03	12.3625	13.05	13.6775	14.83
malic_acid	178	2.3363	1.1171	0.74	1.6025	1.865	3.0825	5.8
ash	178	2.3665	0.2743	1.36	2.21	2.36	2.5575	3.23
alcalinity_of_ash	178	19.4949	3.3396	10.6	17.2	19.5	21.5	30
magnesium	178	99.7416	14.2825	70	88	98	107	162
total_phenols	178	2.2951	0.6259	0.98	1.7425	2.355	2.8	3.88
flavanoids	178	2.0293	0.9989	0.34	1.205	2.135	2.875	5.08
nonflavanoid_phenols	178	0.3619	0.1245	0.13	0.27	0.34	0.4375	0.66
proanthocyanins	178	1.5909	0.5724	0.41	1.25	1.555	1.95	3.58
color_intensity	178	5.0581	2.3183	1.28	3.22	4.69	6.2	13
hue	178	0.9574	0.2286	0.48	0.7825	0.965	1.12	1.71
od280/od315_of_dilute	178	2.6117	0.71	1.27	1.9375	2.78	3.17	4
proline	178	746.8933	314.9075	278	500.5	673.5	985	1680

Średnia zawartość alkoholu to 13%. Wina w tym zbiorze mają od 11% do prawie 15% alkoholu. Połowa win ma zawartość alkoholu poniżej 12.96%**.**
Prolina i Kwas jabłkowy mają rozkład prawoskośny. Możemy to wywnioskować, porównując średnią z medianą. Dla proline średnia (746.9) jest znacznie wyższa niż mediana (500.5). Widzimy również wysokie odchylenie standardowe dla proliny.
Dla malic_acid: średnia (2.33) jest również wyższa niż mediana (1.60). Oznacza to, że większość win ma stosunkowo niską zawartość proliny i kwasu jabłkowego, ale istnieje niewielka grupa win z wysokimi wartościami.
Widzimy również, że wina mocno różnią się barwą, o czym świadczy wysokie odchylenie standardowe, oraz spore różnice między min/max, a średnią/medianą.

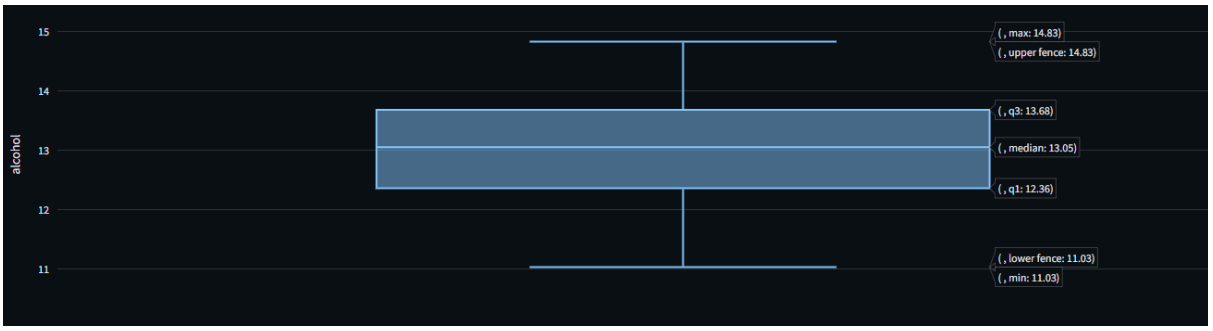
Korelacja Pearsona

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline
alcohol	1	0.0944	0.2115	-0.3102	0.2708	0.2891	0.2368	-0.1559	0.1367	0.5464	-0.0717	0.0723	0.6437
malic_acid	0.0944	1	0.164	0.2885	-0.0546	-0.3352	-0.411	0.293	-0.2207	0.249	-0.5613	-0.3687	-0.192
ash	0.2115	0.164	1	0.4434	0.2866	0.129	0.1151	0.1862	0.0097	0.2589	-0.0747	0.0039	0.2236
alcalinity_of_ash	-0.3102	0.2885	0.4434	1	-0.0833	-0.3211	-0.3514	0.3619	-0.1973	0.0187	-0.274	-0.2768	-0.4406
magnesium	0.2708	-0.0546	0.2866	-0.0833	1	0.2144	0.1958	-0.2563	0.2364	0.2	0.0554	0.066	0.3934
total_phenols	0.2891	-0.3352	0.129	-0.3211	0.2144	1	0.8646	-0.4499	0.6124	-0.0551	0.4337	0.6999	0.4981
flavanoids	0.2368	-0.411	0.1151	-0.3514	0.1958	0.8646	1	-0.5379	0.6527	-0.1724	0.5435	0.7872	0.4942
nonflavanoid_phenols	-0.1559	0.293	0.1862	0.3619	-0.2563	-0.4499	-0.5379	1	-0.3658	0.1391	-0.2626	-0.5033	-0.3114
proanthocyanins	0.1367	-0.2207	0.0097	-0.1973	0.2364	0.6124	0.6527	-0.3658	1	-0.0252	0.2955	0.5191	0.3304
color_intensity	0.5464	0.249	0.2589	0.0187	0.2	-0.0551	-0.1724	0.1391	-0.0252	1	-0.5218	-0.4288	0.3161
hue	-0.0717	-0.5613	-0.0747	-0.274	0.0554	0.4337	0.5435	-0.2626	0.2955	-0.5218	1	0.5655	0.2362
od280/od315_of_dilute	0.0723	-0.3687	0.0039	-0.2768	0.066	0.6999	0.7872	-0.5033	0.5191	-0.4288	0.5655	1	0.3128
proline	0.6437	-0.192	0.2236	-0.4406	0.3934	0.4981	0.4942	-0.3114	0.3304	0.3161	0.2362	0.3128	1

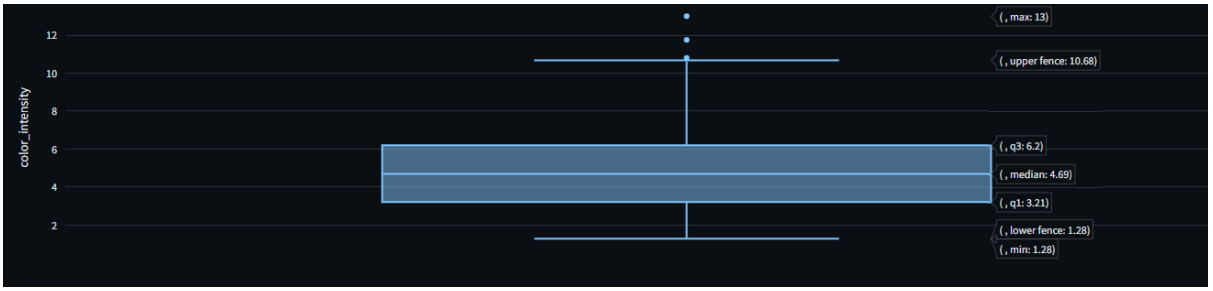
Najbardziej rzucająca się w oczy jest wysoka korelacja dodatnia między flavanoids a total_phenols. To logiczne z chemicznego punktu widzenia. Flawonoidy to po prostu jeden z typów fenoli.
Mocniejsze wina mają tendencję do posiadania intensywniejszej barwy oraz zawierają wyższe stężenia proliny.

Wykresy pudełkowe

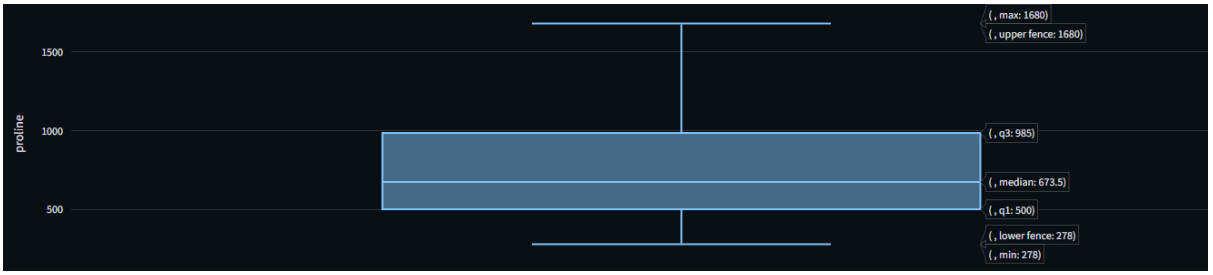
Alkohol



Intesywność koloru

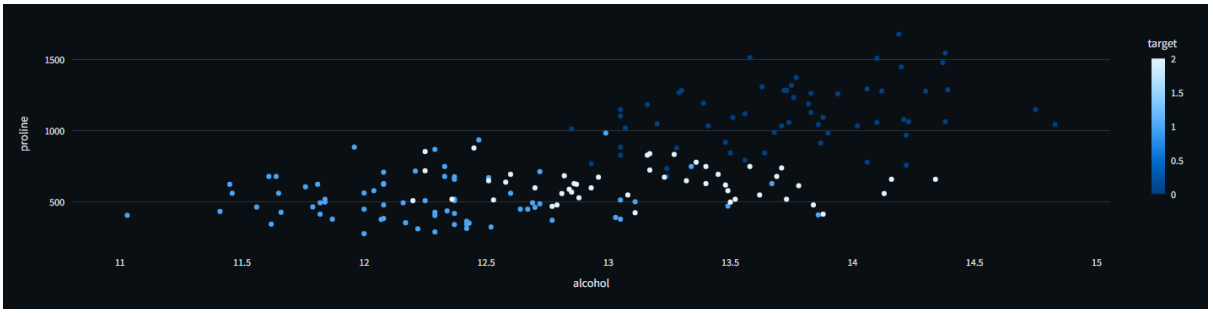


Proliny



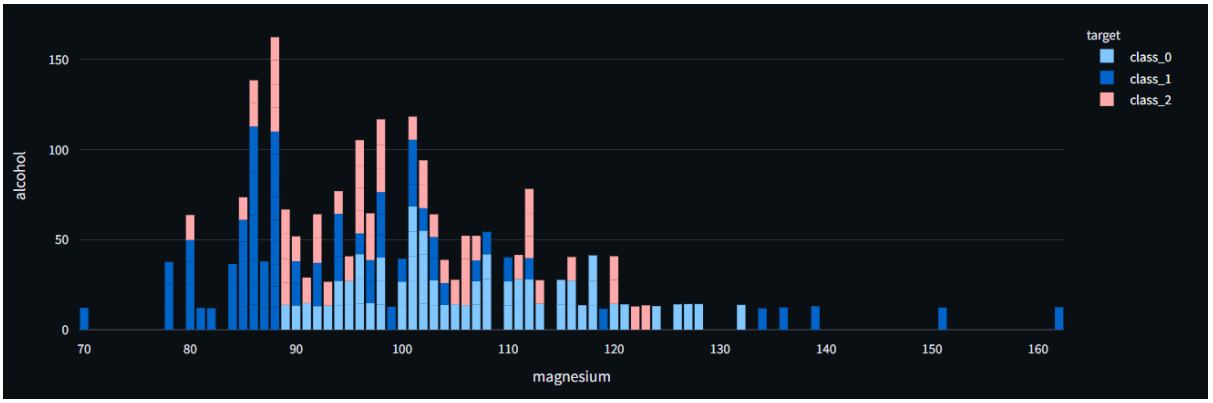
Wykres punktowy

alkohol/proliny

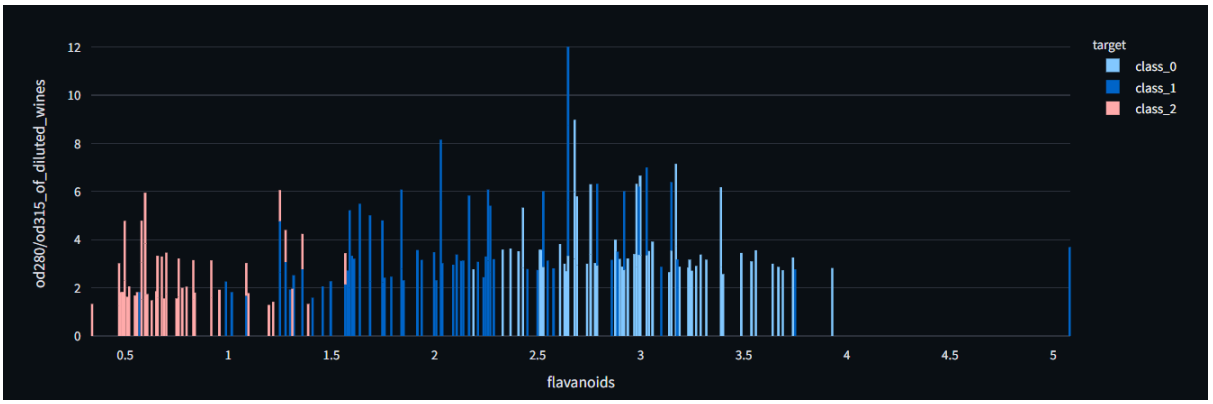


Wykres słupkowy

Magnez/alkohol



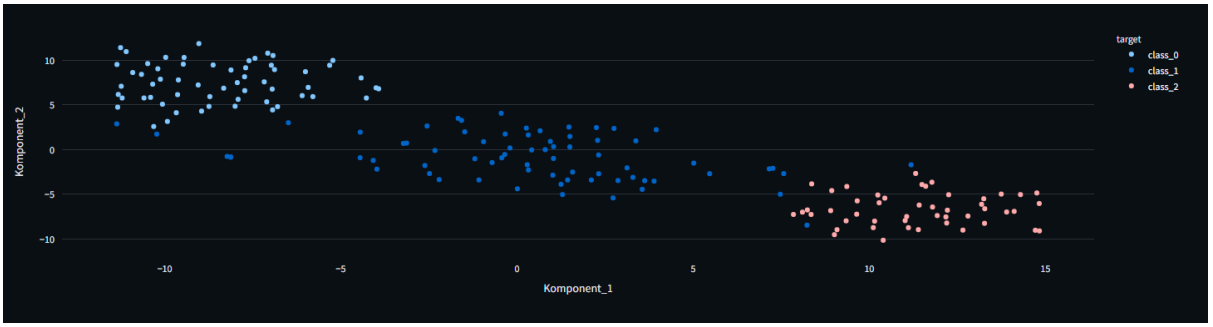
flawonoidy/absorbancja



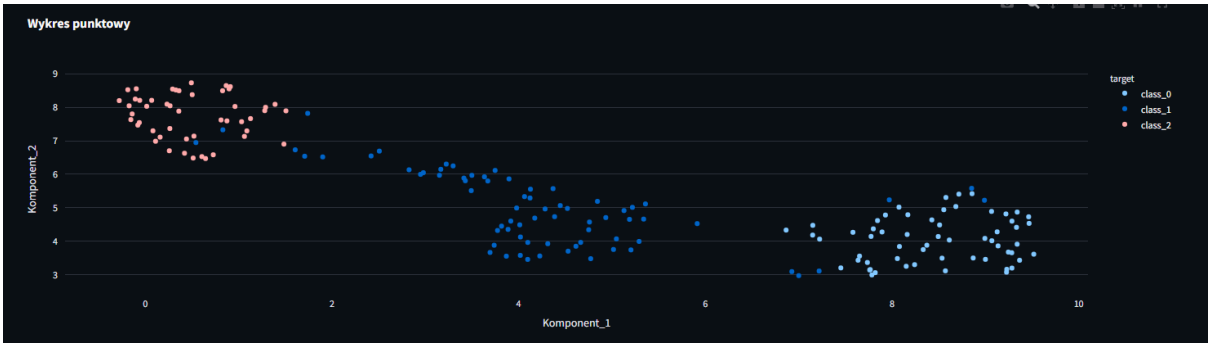
Liczba flawonoidów bardzo dobrze odróżnia klasy win uwzględnionych w naszej próbce. Wino typu 2 ma mało flawonoidów oraz ma niski stosunek absorbancji.

Redukcja wymiarowości

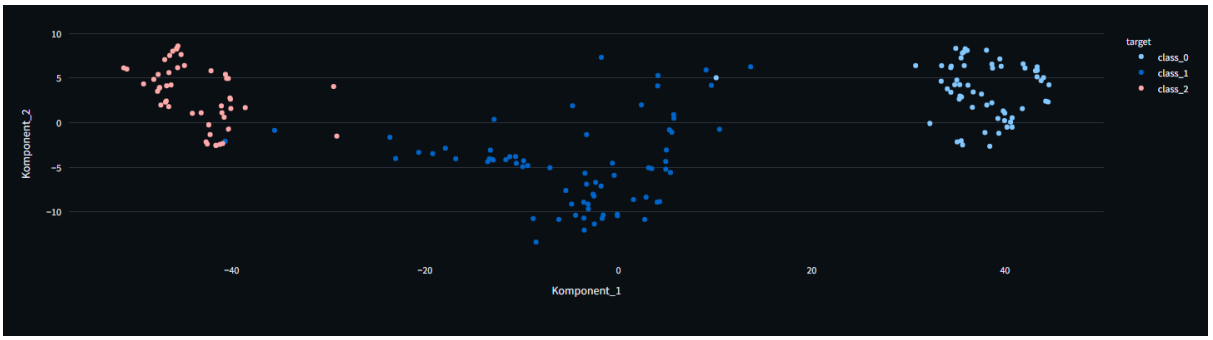
t-sne



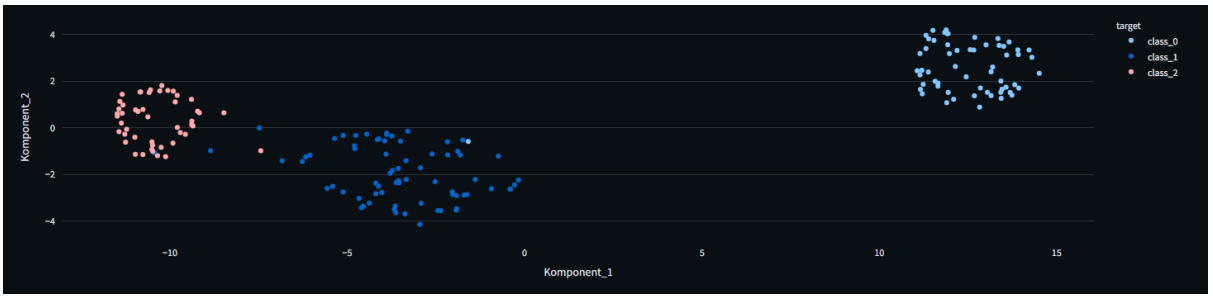
UMAP



TRIMAP



PacMAP



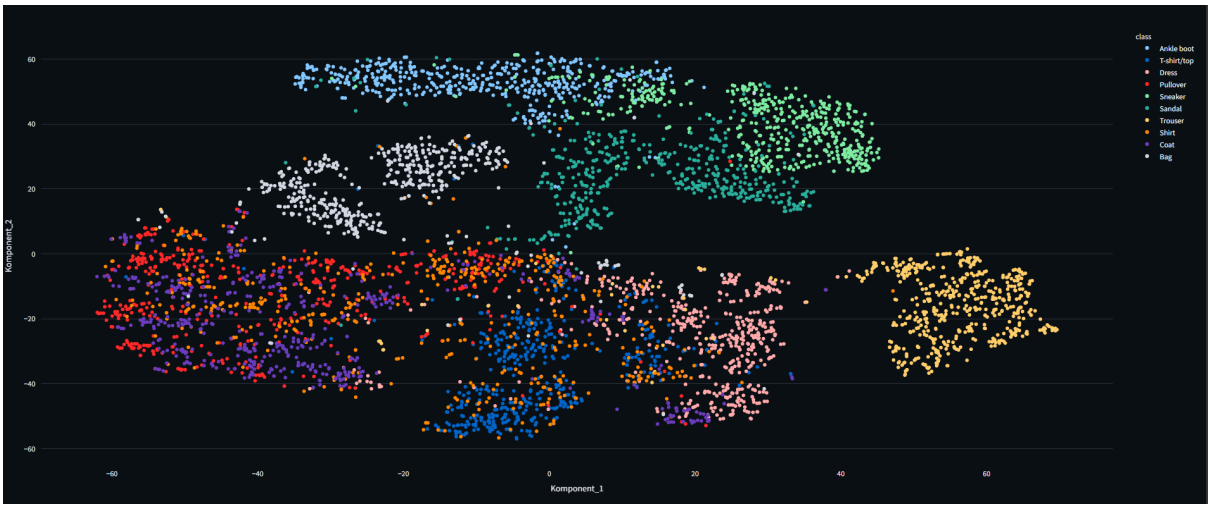
Wszystkie algorytmy redukcji wymiarowości oddzieliły od siebie klasy win. Najlepiej poradził sobie PacMap, ponieważ stworzył on najbardziej zwarte i jednocześnie najdalej od siebie oddalone klastry dla każdej z trzech klas.

Zbiór danych FMNIST

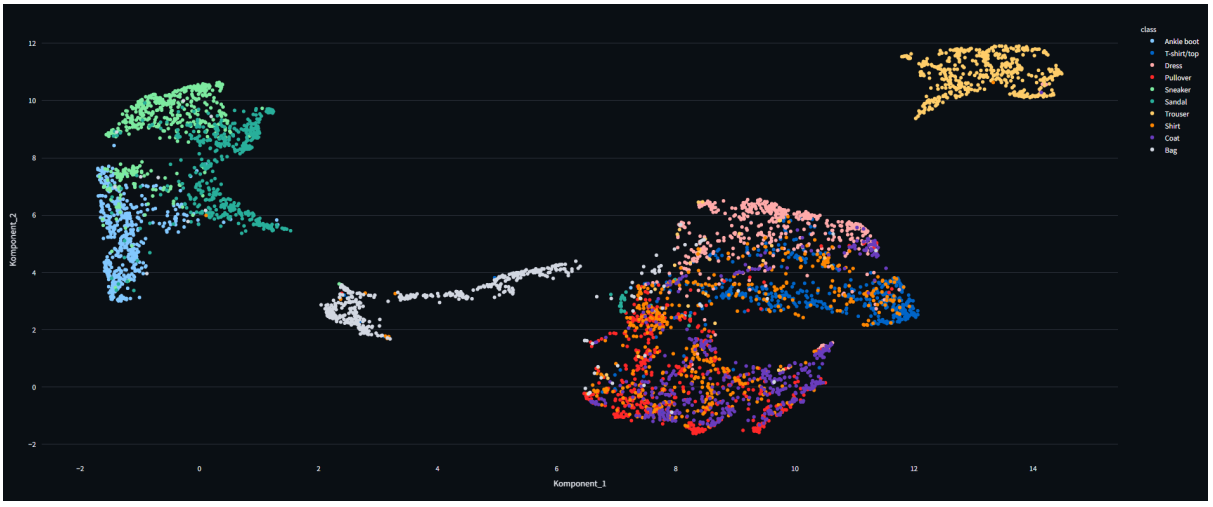
Fashion-MNIST to zbiór danych składający się z obrazów artykułów odzieżowych. Został stworzony przez firmę Zalando Research jako bezpośredni, nowocześniejszy zamiennik dla klasycznego zbioru danych MNIST, który zawiera odręcznie pisane cyfry.

Z racji tego, że analiza danych tutaj nie ma większego sensu od razu przejdziemy do redukcji wymiarowości.

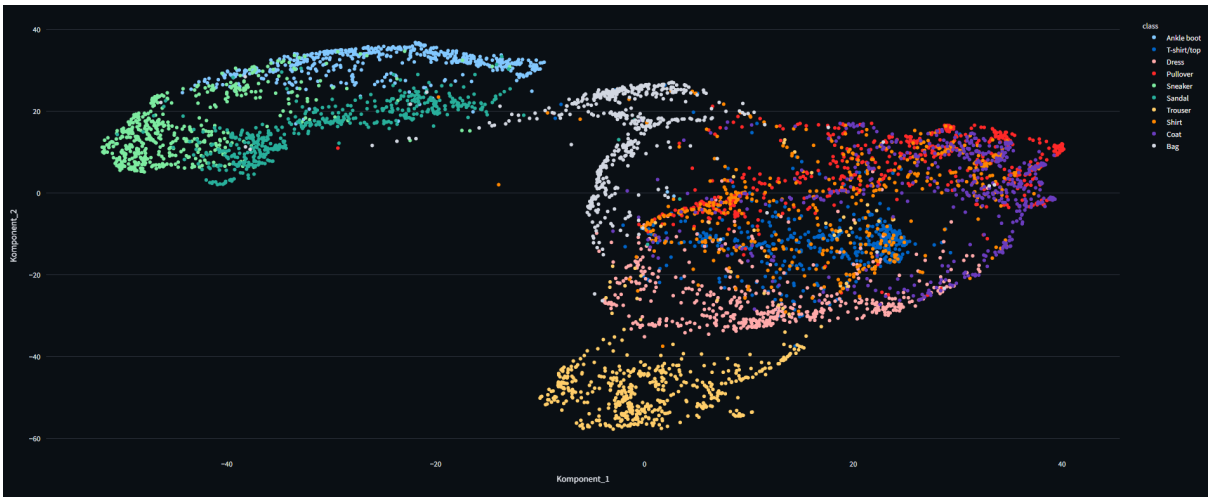
t-sne



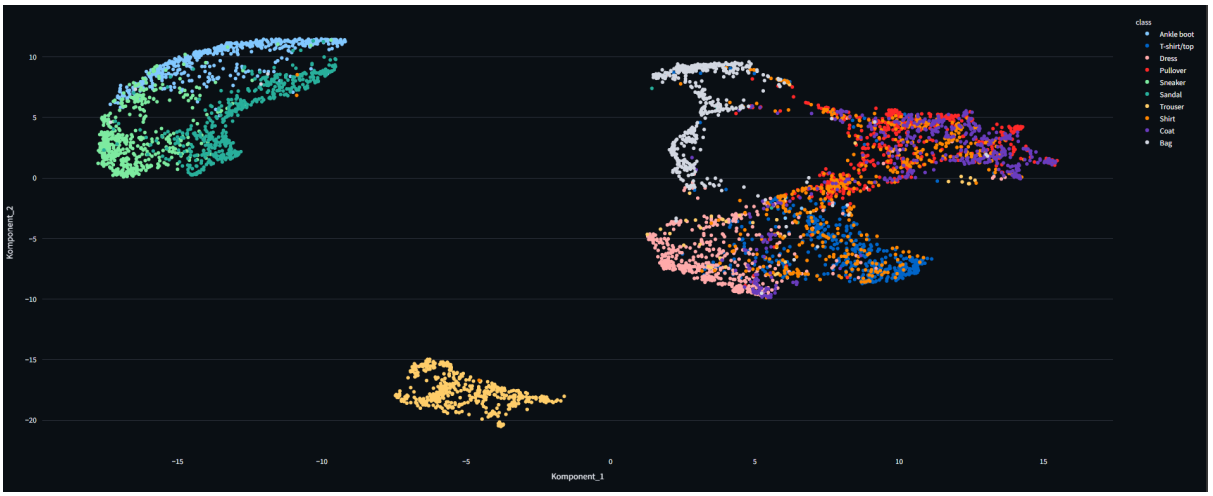
UMAP



TriMap



PacMap



Wszystkie cztery zastosowane algorytmy (t-SNE, UMAP, TriMap, PacMap) z powodzeniem przedstawiły wewnętrzną strukturę zbioru danych. Naszym zdaniem najlepiej wypadł UMAP i PACMAP, gdyż globalnie odseparowały od siebie klastry, zachowując większą odległość, tworząc spójny obraz.