

Wydobywanie wiedzy ze środowiskowych danych przestrzennych - analiza zbioru Covertype

Maciej Dąbrowski 223349 Jakub Dąbrowski 223444
Oleksandr Babenkov 223387 Lizaveta Brazinskaya 223375

8 kwietnia 2025

Streszczenie

W niniejszym projekcie analizujemy zbiór danych Covertype w celu znalezienia wzorców w typach pokrycia terenu. Wykorzystujemy metody grupowania i klasyfikacji, takie jak k-srednich, BIRCH oraz DBSCAN, aby zidentyfikować struktury w danych oraz najlepsze zmienne numeryczne do predykcji. Dzięki analizie statystycznej i eksploracyjnej określamy kluczowe cechy, które mają największy wpływ na podział danych na klastry.

W niniejszej pracy analizowane są dane Covertype dostępne w repozytorium UCI. Wyniki porównano w celu oceny jakości różnych podejść.

1 Wprowadzenie

1.1 Słowa kluczowe

K-means, DBSCAN, BIRCH, Covertype, Classification, Clusters, Forest, Elevation, Slope, Aspect

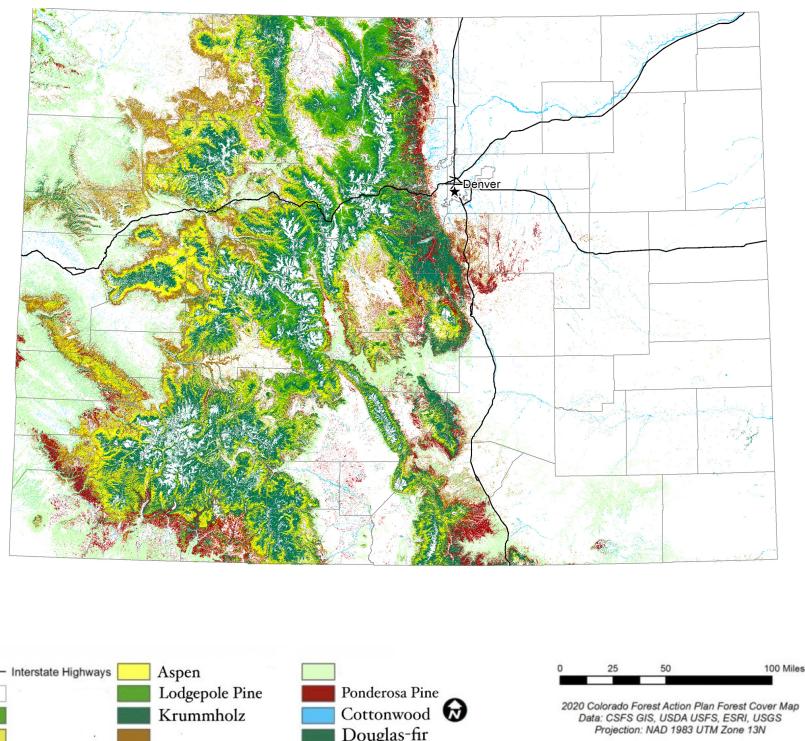
1.2 Typy pokrycia terenu

Dane ze zbioru Covertype zawierają informacje o różnych typach pokrycia terenu w oparciu o atrybuty geograficzne i ekologiczne, lista w kolejności ze zbioru (1-7) przedstawiona poniżej wraz ciekawostkami:

- Spruce/Fir – Świerk/Jodła – Dominuje w chłodnym klimacie górkim, a świerki mają płytke korzenie, przez co łatwo się przewracają.
- Lodgepole Pine – Sosna wydmowa – Jej szyszki otwierają się po pożarach, co pozwala jej szybko kolonizować spalone tereny.
- Ponderosa Pine – Sosna żółta – Osiąga ponad 70 m wysokości, a jej kora pachnie wanilią lub masłem.
- Cottonwood/Willow – Topola/Wierzba – Szybko rosnące drzewa nadrzeczne, a wierzby zawierają naturalny składnik aspiryny.

- Aspen– Osika– Tworzy największy organizm na świecie dzięki wspólnemu systemowi korzeni
- Douglas-fir – Jedlica Douglasta – Nie jest prawdziwą jodłą i może dorastać do 90 m, będąc cennym surowcem drzewnym.
- Krummholtz – Karłowate drzewa górskie – Rosną w ekstremalnych warunkach górskich, przybierając zdeformowane, karłowane kształty.

Dodatkowo przedstawiono wizualizację w postaci mapy (Rys. 1).



Rysunek 1: Mapa lasów



Rysunek 2: Krummholtz



Rysunek 3: Sosna wydmowa



Rysunek 4: Świerk/jodła

2 Przedmiot badania

2.1 Cel i zakres badania

Naszym celem jest analiza i modelowanie danych ze zbioru Covertype w celu znalezienia najciekawszych wzorców oraz identyfikacji kluczowych cech do klasyfikacji/predykcji typów pokrycia terenu. By ograniczyć ilość obliczeń i złożoność wykresu, wybraliśmy losowo 5% zbioru, metoda `.sample()`

2.2 Przegląd literatury

Zbiór Covertype był szeroko analizowany w kontekście klasyfikacji i grupowania. Blac-kard i Dean [1] porównali różne metody klasyfikacyjne, wskazując na wysoką skuteczność modeli opartych na uczeniu maszynowym. Jain [3] omówił techniki grupowania, podkre-ślając, że metoda KMeans sprawdza się w przypadku zbiorów o strukturze sferycznej, natomiast podejścia hierarchiczne, takie jak BIRCH, są efektywne dla dużych zbiorów danych. Ester i in. [4] wprowadzili algorytm DBSCAN, który umożliwia identyfikację nieregularnych klastrów oraz eliminację szumu w danych.

Literatura wskazuje, że zastosowane metody – KMeans, BIRCH i DBSCAN – pozwa-lają na skutecną analizę struktury danych Covertype, w zależności od ich rozkładu i charakterystyki.

2.3 Opis danych i zmiennych

Dane Covertype składają się z 581012 obserwacji i 54 cech, w tym jednej zmiennej celu określającej typ pokrycia terenu. Dla każdej metody dane były standaryzowane za po-mocą biblioteki python `sklearn.preprocessing - StandardScaler`. Były wykorzystane nu-meryczne zmienne, a binarne nie. Z 10 numerycznych zmiennych najbardziej znaczace : **Elevation**(Wysokość), **Slope** (Nachylenie), **Aspect** (Orientacja stoku). Cechy można po-dzielić na następujące kategorie:

- **Cechy topograficzne:**
 - **Elevation** – wysokość nad poziomem morza [m].
 - **Aspect** – orientacja stoku w stopniach (0–360).
 - **Slope** – nachylenie stoku w stopniach.
- **Odległość od hydrologii:**
 - **Horizontal Distance to Hydrology** – odległość w poziomie do najbliższego cieku wodnego [m].
 - **Vertical Distance to Hydrology** – odległość w pionie do najbliższego cieku wodnego [m].
- **Odległość od dróg i miejscowości:**
 - **Horizontal Distance to Roadways** – odległość do najbliższej drogi [m].
 - **Hillshade 9am, Hillshade 3pm, Hillshade noon** – ilość światła słonecz-nego o określonych godzinach.

- **Horizontal Distance to Fire Points** – odległość do najbliższego punktu zarejestrowanego pożaru [m].

- **Rodzaj gleby:**

- Zbiór danych zawiera 40 binarnych zmiennych reprezentujących typy gleb. Każda z tych zmiennych przyjmuje wartość 0 lub 1, wskazując na nieobecność lub obecność danego typu gleby w danej próbce. Przykładowo, jeśli zmienna **Soil_Type1** ma wartość 1, oznacza to, że próbka pochodzi z obszaru o typie gleby 1; jeśli ma wartość 0, typ gleby 1 jest nieobecny.

- **Typ pokrycia terenu (zmienna docelowa):**

- Zmienna kategoryczna (1–7) odpowiadająca typowi pokrycia terenu.

- **Brakujące dane (null oraz odstające dane):**

- Odstające dane zostały usunięty przy pomocy programu `odstajace.py`. Natomiast czyste dane zostały zapisane w pliku `dataset.csv`.
- Wartość null nie występowała w oryginalnym zbiorze danych, dlatego nie było potrzeby jej uzupełniania.

Podstawowe statystyki opisowe przedstawiono w tabelach 1 2.

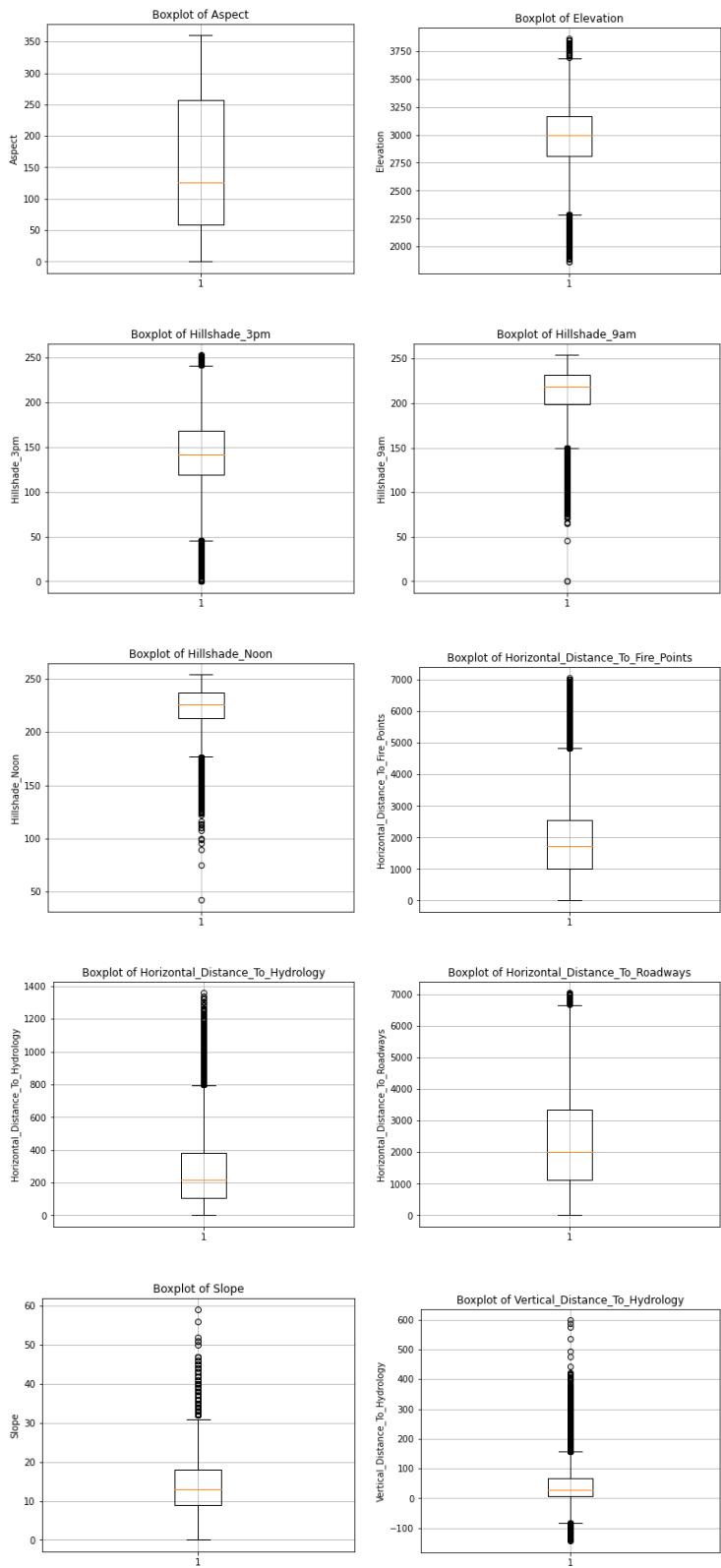
Zmienna	Mean	Median	Mode	Variance	Min	Max	Range
Wysokość terenu (Elevation)	2960.47	2998.0	3014	77463.04	1861	3858	1997
Kierunek nachylenia stoku (Aspect)	154.84	126.0	45	12391.14	0	360	360
Nachylenie terenu (Slope)	14.09	13.0	12	55.95	0	59	59
Odległość pozioma do hydrologii (Horizontal Distance to Hydrology)	269.03	218.0	30	44971.64	0	1359	1359
Odległość pionowa do hydrologii (Vertical Distance to Hydrology)	45.96	29.0	0	3359.46	-144	598	742
Odległość pozioma do dróg (Horizontal Distance to Roadways)	2358.16	2002.0	150	2443146.04	0	7038	7038
Oświetlenie terenu o 9:00 (Hillshade 9am)	212.47	218.0	226	710.33	0	254	254
Oświetlenie terenu o 12:00 (Hillshade Noon)	223.27	226.0	228	395.19	42	254	212
Oświetlenie terenu o 15:00 (Hillshade 3pm)	142.10	142.0	145	1460.69	0	253	253
Odległość pozioma do punktów przeciwożarowych (Horizontal Distance to Fire Points)	1984.23	1714.0	618	1769210.21	0	7050	7050

Tabela 1: Podstawowe statystyki dla zmiennych terenu – część 1

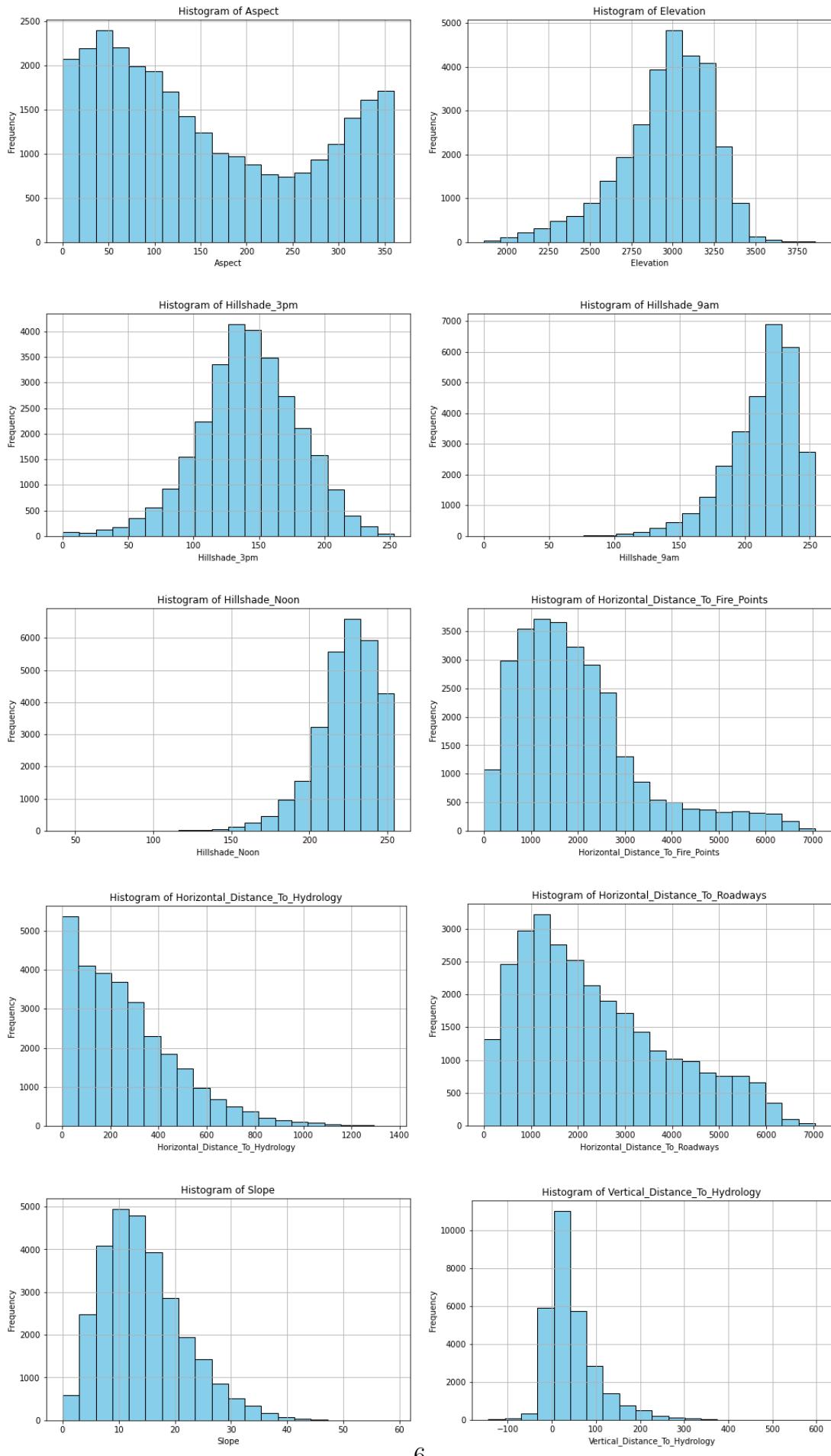
Zmienna	Standard Deviation	IQR	Skośność	Współczynnik zmienności
Wysokość terenu (Elevation)	278.32	350.0	-0.81	0.0940
Kierunek nachylenia stoku (Aspect)	111.32	198.0	0.42	0.7189
Nachylenie terenu (Slope)	7.48	9.0	0.80	0.5309
Odległość pozioma do hydrologii (Horizontal Distance to Hydrology)	212.07	276.0	1.14	0.7883
Odległość pionowa do hydrologii (Vertical Distance to Hydrology)	57.96	60.0	1.84	1.2612
Odległość pozioma do dróg (Horizontal Distance to Roadways)	1563.06	2222.0	0.71	0.6628
Oświetlenie terenu o 9:00 (Hillshade 9am)	26.65	33.0	-1.21	0.1254
Oświetlenie terenu o 12:00 (Hillshade Noon)	19.88	24.0	-1.11	0.0890
Oświetlenie terenu o 15:00 (Hillshade 3pm)	38.22	49.0	-0.27	0.2690
Odległość pozioma do punktów przeciwożarowych (Horizontal Distance to Fire Points)	1330.12	1527.0	1.28	0.6703

Tabela 2: Podstawowe statystyki dla zmiennych terenu – część 2

Dodatkowo przedstawiono wizualizację w postaci wykresów pułapkowych (Rys. 5) Dodatkowo przedstawiono wizualizację w postaci histogramów (Rys. 6).



Rysunek 5: Wykresy pułapkowe dla wybranych zmiennych.



3 Opis metod

3.1 KMeans

Metoda KMeans polega na podziale zbioru danych na ustaloną z góry liczbę klastrów (k), przy czym każdy punkt przypisywany jest do najbliższego centroidu, czyli reprezentanta klasy. W przeciwieństwie do podejścia stosowanego w DBSCAN, KMeans wymaga wcześniejszego określenia liczby klastrów i zakłada, że klastery mają stosunkowo regularny, kulisty kształt. Celem algorytmu jest minimalizacja sumy kwadratów odległości punktów od ich odpowiadających centroidów, co pozwala na wydobycie struktury danych.

Proces działania algorytmu rozpoczyna się od ustalenia liczby klastrów, która ma zostać wyodrębniona. Następnie wybierane są k losowych punktów, pełniących funkcję początkowych centroidów. Każdemu punktowi danych przypisywany jest klasa, której centroid znajduje się najbliżej, najczęściej według odległości euklidesowej. Po dokonaniu tego przypisania dla każdego klastra obliczana jest nowa wartość centroidu jako średnia wartości punktów, które do niego należą. Kolejnym krokiem jest powtarzanie procesu przypisywania punktów oraz aktualizacji centroidów aż do momentu, gdy zmiany w podziale na klastery staną się nieistotne lub zostanie spełniony ustalony warunek zbieżności.

1. Odległość euklidesowa między punktem x_i a centroidem μ_j :

$$d(x_i, \mu_j) = \sqrt{\sum_{m=1}^M (x_{im} - \mu_{jm})^2} \quad (1)$$

2. Aktualizacja położenia centroidów:

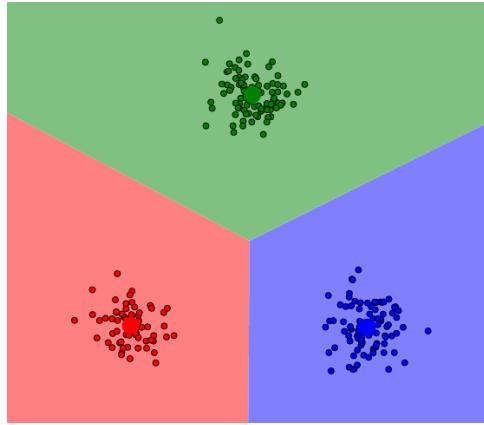
$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

gdzie $|C_j|$ to liczba punktów w klastrze C_j .

Metoda KMeans cechuje się wysoką szybkością oraz wydajnością, co sprawia, że jest szczególnie przydatna przy analizie dużych zbiorów danych. Łatwość implementacji i intuicyjna interpretacja wyników, gdzie centroidy reprezentują „środek ciężkości” klastrów, umożliwiają szybkie zrozumienie struktury danych. Dodatkowym atutem jest jej skalowalność, co pozwala na szerokie zastosowanie w różnych dziedzinach analizy danych.

Należy jednak pamiętać, że metoda ta wiąże się również z pewnymi ograniczeniami. Przede wszystkim wymaga ona wcześniejszego ustalenia liczby klastrów, co może być problematyczne, gdy liczba struktur w danych nie jest oczywista. Losowy wybór początkowych centroidów może prowadzić do zmiennych wyników, a niewłaściwa inicjalizacja często skutkuje zbieżnością do lokalnego optimum. Ponadto, przyjęcie założenia o kulistym kształcie klastrów ogranicza skuteczność algorytmu w przypadku danych o nieregularnych strukturach, a obecność wartości odstających może negatywnie wpływać na pozycje centroidów.

Podsumowując, metoda KMeans stanowi efektywne i skalowalne narzędzie do klasteryzacji danych, szczególnie gdy struktura zbioru odpowiada przyjętym założeniom. Jej główne zalety, takie jak szybkość działania, prostota implementacji i interpretacji wyników, są równoważone przez konieczność wcześniejszego określenia liczby klastrów oraz wrażliwość na inicjalizację i wartości odstające.



Rysunek 7: Wizualizacja klastrów KMeans z trzema grupami.

3.2 BIRCH

BIRCH to metoda grupowania danych, która umożliwia tworzenie klastrów na różnych poziomach szczegółowości, co pozwala na analizowanie danych w formie drzewa (dendrogramu) i elastyczną interpretację ich struktury bez konieczności określania liczby klastrów z góry. Metoda ta wykorzystuje dwa podejścia do organizacji danych. W podejściu agresywnym (bottom-up, agglomerative) każdy punkt początkowo traktowany jest jako oddzielny klaster, a następnie punkty te są łączone w większe grupy na podstawie ich wzajemnego podobieństwa. Natomiast w podejściu dzielącym (top-down, divisive) wszystkie punkty tworzą jeden duży klaster, który stopniowo jest rozbijany na mniejsze podgrupy. Takie rozwiązanie pozwala na analizę danych na różnych poziomach szczegółowości, co jest szczególnie przydatne przy wyszukiwaniu naturalnych struktur w zbiorze danych.

1. Definicja wektora cech grupowania (CF) dla klastra C_j :

$$CF_j = \left(N_j, \sum_{i=1}^{N_j} x_i, \sum_{i=1}^{N_j} x_i^2 \right) \quad (3)$$

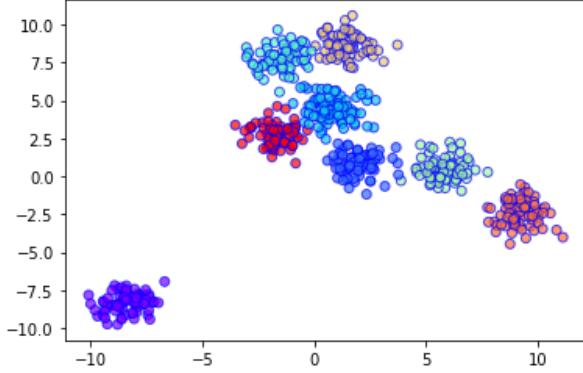
gdzie: - N_j to liczba punktów w klastrze, - $\sum_{i=1}^{N_j} x_i$ to suma wektorów punktów, - $\sum_{i=1}^{N_j} x_i^2$ to suma kwadratów współrzędnych punktów.

2. Odległość między dwoma klastrami C_j i C_k można oszacować jako:

$$D(C_j, C_k) = \left\| \frac{1}{N_j} \sum_{i=1}^{N_j} x_i - \frac{1}{N_k} \sum_{i=1}^{N_k} x_i \right\| \quad (4)$$

Zaletą metody BIRCH jest przede wszystkim to, że nie trzeba uprzednio określić liczby klastrów – dendrogram umożliwia znalezienie optymalnego podziału danych. Ponadto, możliwość tworzenia hierarchii klastrów pozwala na dogłębną analizę na wielu poziomach szczegółowości, a metoda ta cechuje się dobrą efektywnością przy analizie mniejszych zbiorów danych. Z drugiej strony, istnieją pewne ograniczenia tej techniki. Dla dużych zbiorów danych BIRCH może działać bardzo wolno, gdyż złożoność obliczeniowa algorytmu wynosi $O(n^2)$ lub $O(n^3)$. Dodatkowo, algorytm jest szczególnie wrażliwy na obecność wartości odstających oraz szumu, co może wpływać na nieoptimalne działanie, a raz utworzone połączenia między klastrami nie mogą być cofnięte, co ogranicza elastyczność w dalszej modyfikacji wyników.

Podsumowując, BIRCH stanowi wartościowe narzędzie analityczne dzięki swojej hierarchicznej strukturze pozwalającej na dogłębną analizę danych na różnych poziomach szczegółowości, jednakże warto mieć na uwadze jego ograniczenia związane z wydajnością przy większych zbiorach oraz podatnością na zakłócenia wywołane przez szum i wartości odstające.



Rysunek 8: Wizualizacja klastrów za pomocą metody BIRCH.

3.3 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to metoda klasteryzacji, która opiera się na analizie gęstości rozmieszczenia punktów w przestrzeni. Algorytm ten grupuje dane na podstawie zagęszczenia – punkty znajdujące się blisko siebie są scalane w jeden klaster, natomiast te pojedyncze lub oddalone od gęstych skupisk traktowane są jako szum. W przeciwieństwie do popularnych metod, takich jak K-Means, DBSCAN nie wymaga wcześniejszego określania liczby klastrów, co pozwala na lepszą eksplorację danych o nieregularnych kształtach i zmiennym rozkładzie.

Proces działania DBSCAN rozpoczyna się od losowego wyboru punktu, który nie został jeszcze przypisany do żadnego klastra. Następnie badane jest jego otoczenie określone przez parametr ε – jeżeli w tym promieniu znajduje się co najmniej określona liczba punktów $MinPts$, punkt ten zostaje uznany za rdzeń nowego klastra. Kolejne punkty z otoczenia są dołączane do klastra, a algorytm rekurencyjnie sprawdza ich sąsiedztwo, co prowadzi do rozrastania się skupiska. Jeśli dany punkt nie spełnia warunków gęstości, jest traktowany jako szum.

1. Warunek przynależności punktu x_i do klastra:

$$|\{x_j \mid d(x_i, x_j) \leq \varepsilon\}| \geq MinPts \quad (5)$$

gdzie $MinPts$ to minimalna liczba sąsiadów w promieniu ε .

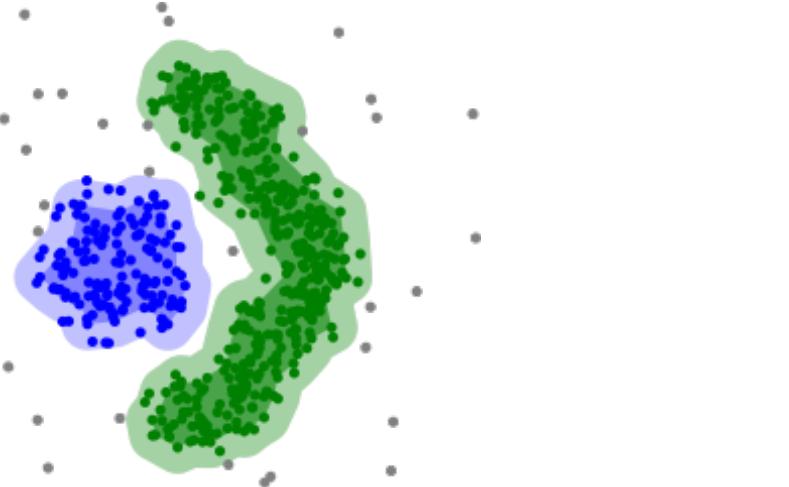
2. Odległość euklidesowa między dwoma punktami:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^M (x_{im} - x_{jm})^2} \quad (6)$$

Do głównych zalet DBSCAN należy fakt, że metoda ta nie wymaga ustalania liczby klastrów z góry, co ułatwia analizę danych, których liczba naturalnych grup nie jest znana. Ponadto, DBSCAN potrafi identyfikować klastry o dowolnych kształtach i rozmaitym zagęszczeniu, a także wykazuje wysoką odporność na szum oraz wartości odstające. Z

drugiej strony, skuteczność algorytmu zależy w dużej mierze od odpowiedniego doboru parametrów ε i $MinPts$, co może stanowić wyzwanie. Ponadto, w przypadku danych o nierównomiernym zagęszczeniu klasteryzacja może być utrudniona, gdyż skupiska o różnej gęstości mogą być nieprawidłowo wydzielane.

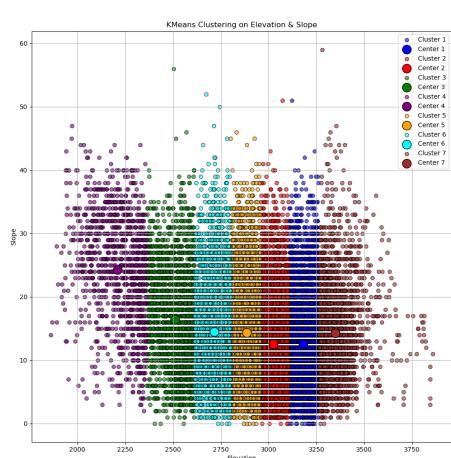
Podsumowując, DBSCAN to efektywna metoda klasteryzacji, idealna do analizy danych o skomplikowanej strukturze. Jego zdolność do wykrywania nieregularnych klastrów oraz odporność na szum sprawiają, że jest to wartościowe narzędzie w wielu zastosowaniach, choć wymaga starannego dobrania parametrów dla uzyskania optymalnych wyników.



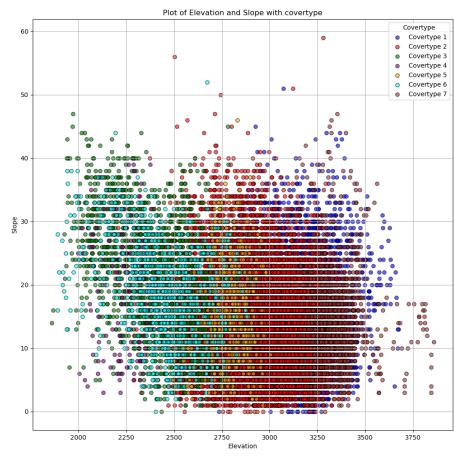
Rysunek 9: Wizualizacja klastrów za pomocą metody DBSCAN.

4 Rezultaty

4.1 Wykresy KMeans

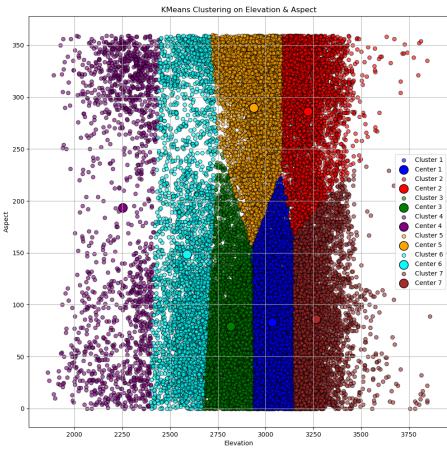


Rysunek 10: Wykres 1.1. Grupowanie KMeans według wysokości i nachylenia

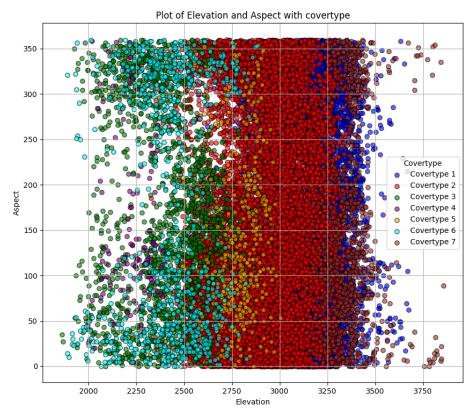


Rysunek 11: Wykres 1.2. Wykres wysokości i nachylenia z typem pokrycia terenu

Pierwszy wykres (KMeans Clustering) pokazuje grupowanie punktów za pomocą algorytmu KMeans, co prowadzi do wyraźnych pionowych pasm, co sugeruje, że wysokość jest dominującą cechą w podziale klastrów. Drugi wykres (Covertype Classification) pokazuje rzeczywisty podział typów pokrycia terenu, który ma bardziej złożony wzór. Widoczne są pewne podobne pionowe pasma, ale klasyfikacja jest bardziej zróżnicowana na różnych poziomach nachylenia. Można zauważyć, że niektóre grupy klastrów w pierwszym wykresie dobrze pokrywają się z rzeczywistymi klasami z drugiego wykresu, ale istnieją obszary, gdzie rzeczywisty podział jest bardziej skomplikowany, co sugeruje użycie większej ilości zmiennych

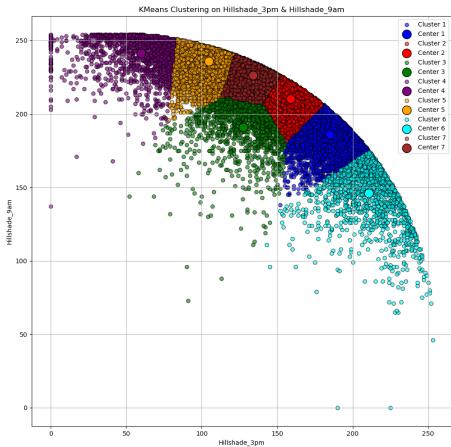


Rysunek 12: Wykres 2.1. Grupowanie KMeans według wysokości i aspektu

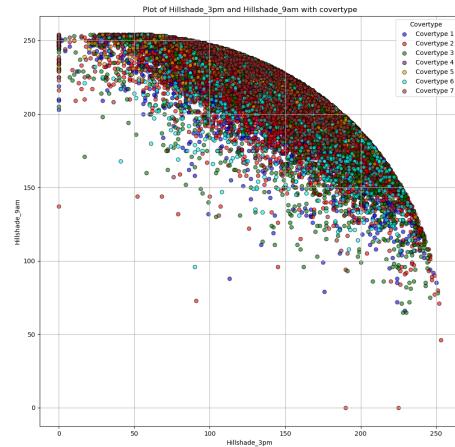


Rysunek 13: Wykres 2.2. Wykres wysokości i aspektu z typem pokrycia terenu

Na pierwszym wykresie widzimy, że klasteryzacja podzieliła dane głównie według wysokości (Elevation) oraz częściowo według aspektu (Aspect). Granice klastrów są wyraźne, co sugeruje, że wysokość ma dominujące znaczenie w procesie grupowania. Rozkład aspektu, czyli orientacji terenu, wydaje się mniej wpływać na podział klastrów, ale w rzeczywistych danych (drugi wykres) widać, że niektóre pokrycia mogą mieć lekką preferencję względem orientacji stoków (np. różne ekspozycje względem słońca mogą wpływać na wegetację). Różnice w pokryciu terenu wynikają prawdopodobnie ze zmian w warunkach siedliskowych, takich jak temperatura, dostępność wody oraz typ gleby. Nizsze wysokości mogą sprzyjać bardziej zróżnicowanej roślinności, podczas gdy na wyższych dominują bardziej odporne gatunki.



Rysunek 14: Wykres 3.1. Grupowanie KMeans według nasłonecznienia o godz. 15 i godz. 9



Rysunek 15: Wykres 3.2. Wykres nasłonecznienia o godz. 15 i godz. 9 z typem pokrycia terenu

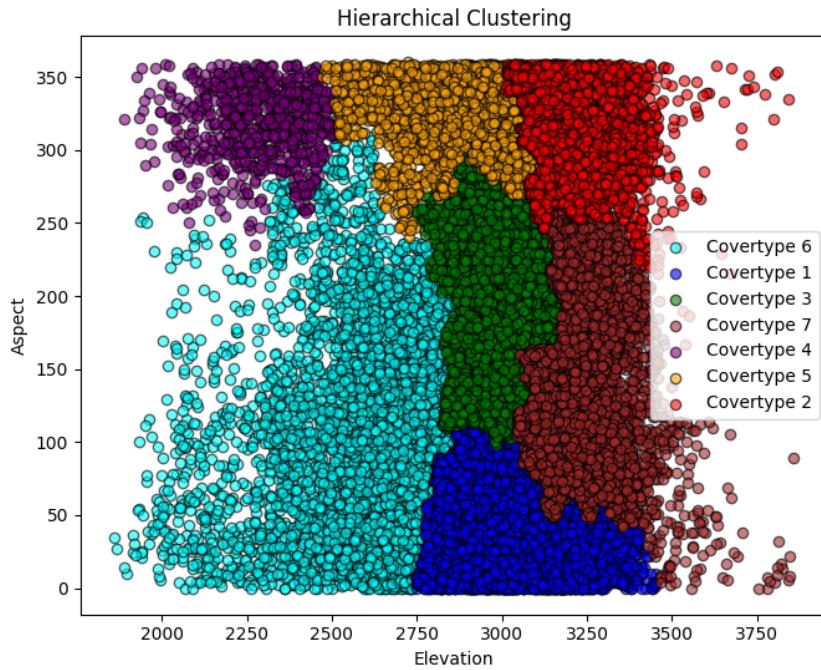
Pierwszy wykres przedstawia wynik klastrowania metodą KMeans na danych dotyczących nasłonecznienia o godz. 15 i godz. 9. Widoczne są wyraźnie zdefiniowane grupy, które układają się w łuk, sugerując, że wartości nasłonecznienia mają ograniczony zakres i mogą być skorelowane. Drugi wykres pokazuje rzeczywiste typy pokrycia terenu (covertype) i widać, że naturalne klasy (covertype) są bardziej wymieszane niż wyniki klastrowania K-Means, co sugeruje, że klasyfikacja rzeczywistych pokryć terenu może być bardziej złożona i nie w pełni uchwytna przez prostą analizę klastrów. Dodatkowo, oba wykresy sugerują, że różne typy pokrycia terenu mają podobne rozkłady w przestrzeni Hillshade 3pm i Hillshade 9am, a krzywa widoczna na obu wykresach przypomina fragment paraboli lub bardziej ogólnie funkcji pierwiastkowej (np. $y = (-x^2 + a^2)^{\frac{1}{2}}$) co sugeruje ograniczenie wartości przez pewien maksymalny promień. Jej kształt wskazuje na związek między Hillshade 3pm i Hillshade 9am, który może wynikać z geometrycznych właściwości nasłonecznienia – np. sposób padania światła o różnych porach dnia na powierzchnię terenu.

4.2 Linki do stron 3D

Poniżej znajdują się wybrane linki:

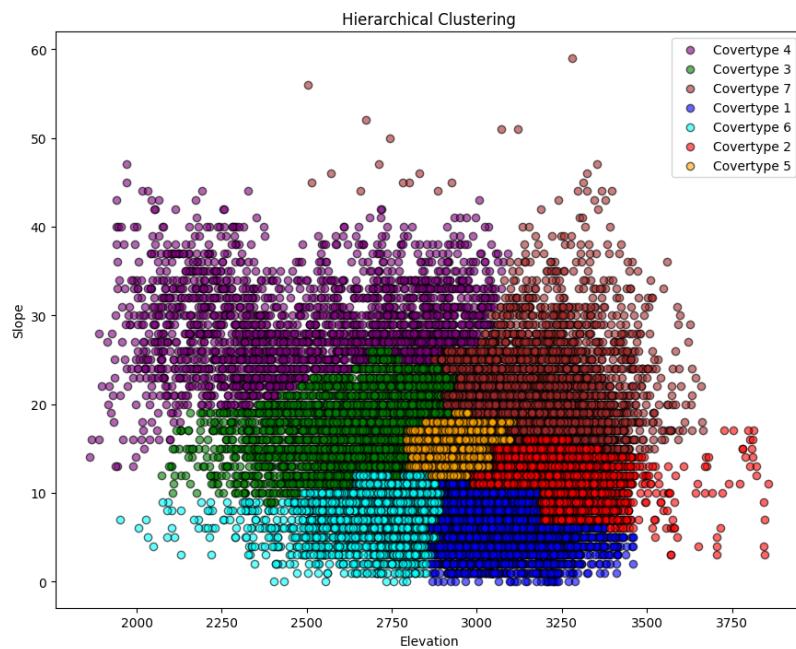
- Link 1 - Cover 3D
- Link 2 - Clusters 3D
- Link 3 - Clusters 3D

4.3 Wykresy Hierarchical - BIRCH



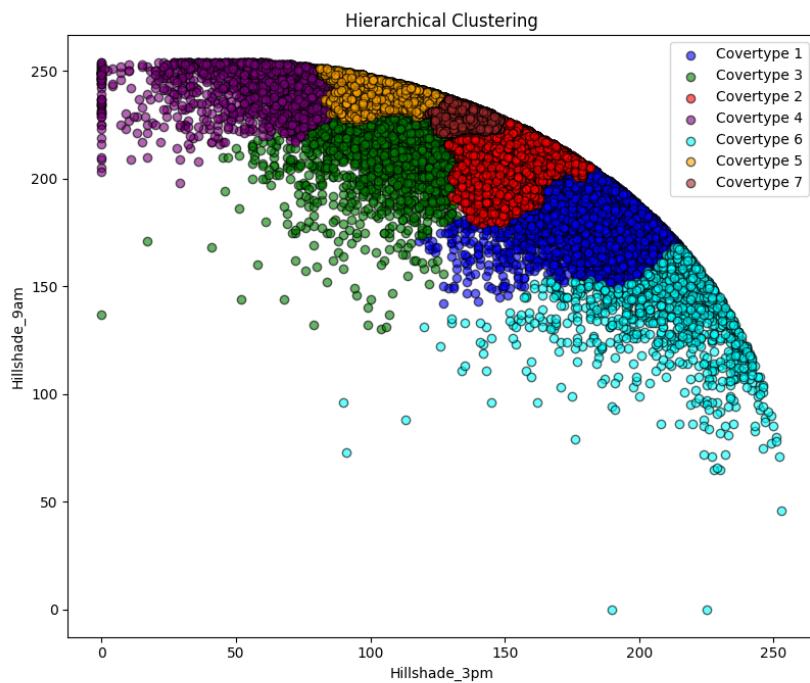
Rysunek 16: Wykres 4.1. Grupowanie BIRCH według wysokości i aspektu z typem pokrycia terenu

Na wykresie widać, że na najwyższych poziomach hierarchii dominującą rolę ponownie odgrywa wysokość (Elevation). Jednakże drzewiasta struktura wyraźnie oddziela także podgrupy, w których istotny jest wpływ parametrów takich jak nachylenie terenu (Slope). Interesujące jest także to, że metoda hierarchiczna pozwala na obserwację naturalnego ciągu podobieństw między danymi, co może ujawniać obszary, gdzie granice klastrów są bardziej rozmyte. Obserwujemy bowiem sytuacje, w których obszary o zbliżonych wartościach Elevation, ale różniące się subtelnie nachyleniem, wskazują na istnienie przejściowych stref siedliskowych. Mogą one być szczególnie cenne, gdyż często odzwierciedlają miejsca, gdzie warunki środowiskowe zmieniają się stopniowo – na przykład przejścia między terenami bardziej zalesionymi a otwartymi polanami.



Rysunek 17: Wykres 4.2. Grupowanie BIRCH według wysokości i nachylenia z typem pokrycia terenu

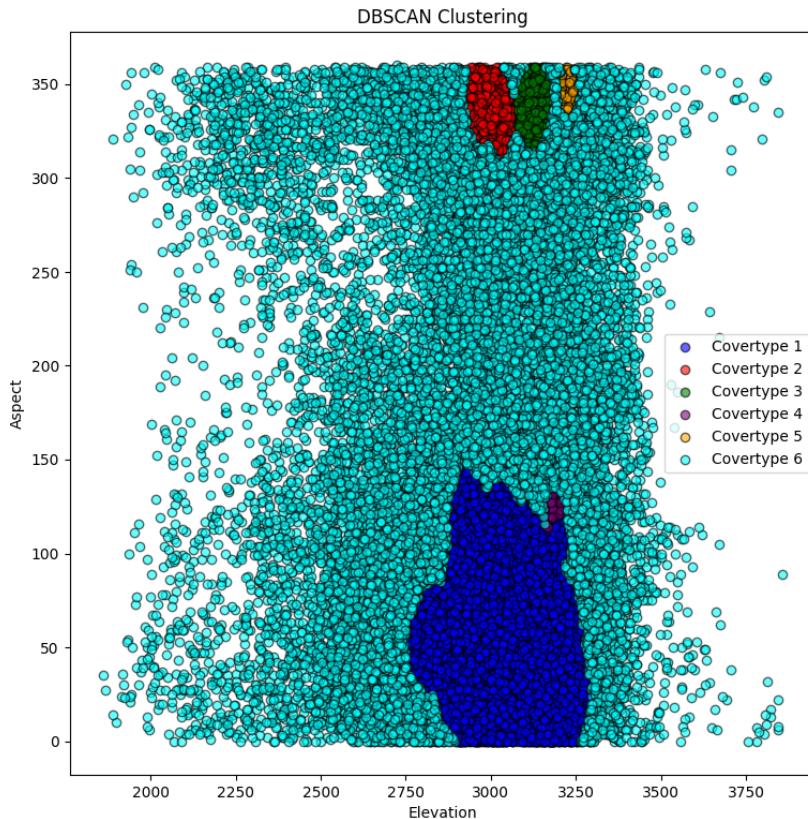
Wykres wskazuje, że wysokość terenu jest nadal kluczowym czynnikiem – punkty odpowiadające różnym typom pokrycia wyraźnie segregują się wzdłuż osi Elevation. Kąt orientacji (Aspect) wnosi dodatkowy wymiar różnicujący – choć nie zawsze decydujący samodzielnie, to jednak w połączeniu z wysokością pozwala na jeszcze dokładniejsze rozróżnienie typów pokrycia. Niestety na wykresach klastry mają inną kolejność (1 w covertype nie jest tą samą 1 w klastrach), wynika to z specyfiki metody i trudności poprawy tego problemu.



Rysunek 18: Wykres 4.3. Grupowanie BIRCH według nasłonecznienia o godz 15. oraz godz. 9 z typem pokrycia terenu

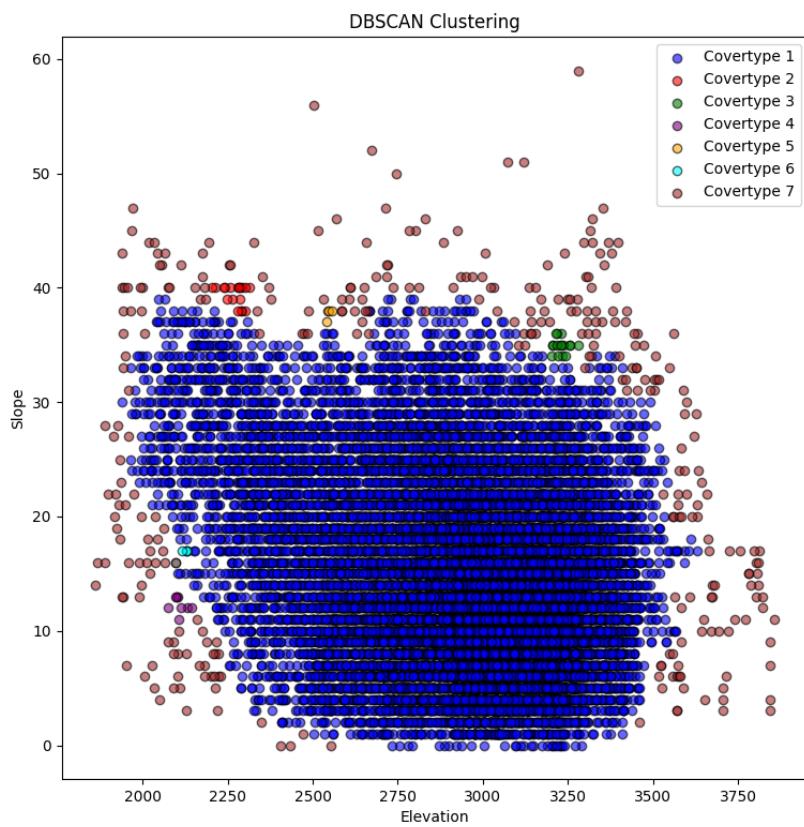
Na wykresie można zauważać wyraźne skupiska punktów, które odpowiadają poszczególnym typom pokrycia. Podział ten sugeruje, że różnice w oświetleniu w różnych porach dnia (9:00 rano vs. 15:00 po południu) są istotnymi czynnikami różnicującymi teren pod względem pokrycia. Każdy typ pokrycia terenu wykazuje charakterystyczny wzorzec rozkładu punktów – na przykład jedna grupa może skupiać się w obszarze o wyższych wartościach Hillshade 3pm i średnich Hillshade 9am, co mogłoby świadczyć o specyficzny ustawieniu względem słońca i unikalnej strukturze roślinności.

4.4 Wykresy DBSCAN



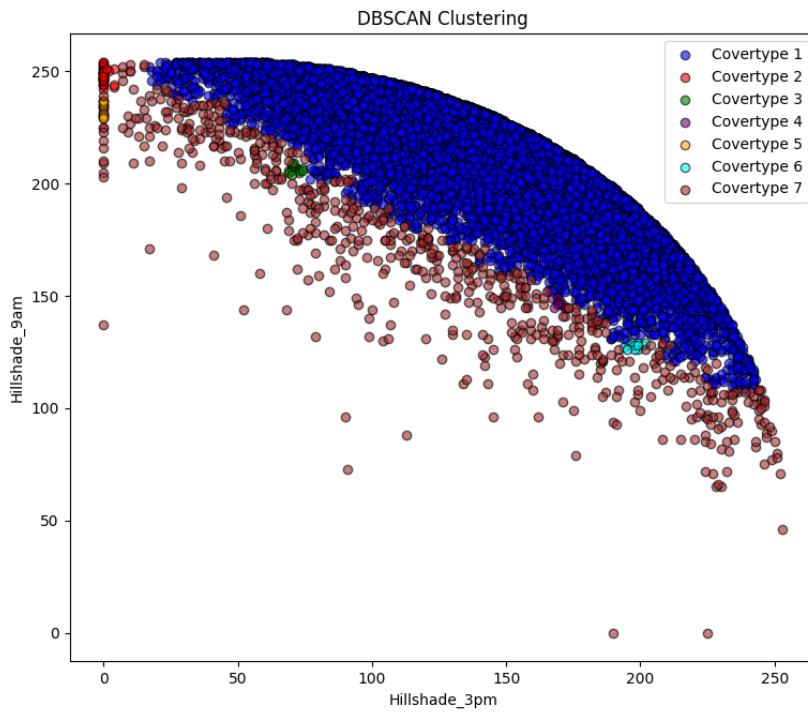
Rysunek 19: Wykres 5.1 Grupowanie DBSCAN według wysokości i aspektu z typem pokrycia terenu

Na przedstawionym wykresie widzimy, że metoda DBSCAN podzieliła dane w sposób odmienny niż KMeans. Kluczową cechą DBSCAN jest zdolność do identyfikacji gęstych skupisk punktów oraz wykrywania punktów odstających (outliers). W tym przypadku klasteryzacja nadal sugeruje silne uzależnienie pokrycia terenu od wysokości (Elevation), ale w przeciwieństwie do K-means, DBSCAN nie wymusza podziału na określona liczbę grup. Można zauważać, że najgęstszy klaster (dominujący kolor, np. niebieski) koncentruje się w dolnym obszarze wykresu, co wskazuje na większą ilość punktów w określonym zakresie wysokości i nachylenia. Ponadto, widać mniejsze, bardziej rozproszone klastry na obrzeżach głównej grupy, co sugeruje, że DBSCAN skutecznie identyfikuje obszary o różnym stopniu gęstości. Może to odzwierciedlać rzeczywiste zróżnicowanie ekosystemów, gdzie niektóre typy pokrycia terenu występują w specyficznych warunkach, a inne mają bardziej rozproszony charakter. Metoda DBSCAN może również skutecznie identyfikować anomalie – pojedyncze punkty lub małe grupy punktów, które nie należą do żadnego większego klastra. Może to wskazywać na rzadkie lub nietypowe obszary, np. miejsca o ekstremalnych warunkach środowiskowych.



Rysunek 20: Wykres 5.2. Grupowanie DBSCAN według wysokości i nachylenia z typem pokrycia terenu

Na przedstawionym wykresie widać, że metoda DBSCAN wykryła kilka wyraźnych skupisk danych na podstawie wysokości (Elevation) oraz aspektu (Aspect). Podobnie jak w przypadku poprzednich metod, wysokość nadal wydaje się dominującym czynnikiem wpływającym na podział klastrów. Można zauważyć, że największy i najbardziej wyraźny klaster znajduje się w dolnej części wykresu (niższe wysokości), co może sugerować, że określony typ pokrycia terenu dominuje na niższych wysokościach. W wyższych partiach pojawiają się mniejsze, ale bardziej skoncentrowane skupiska, co może wynikać ze specyficznych warunków środowiskowych (np. innego typu gleby czy ekspozycji stoków).



Rysunek 21: Wykres 5.3 Grupowanie DBSCAN według nasłonecznienia o 3PM oraz o 9AM z typem pokrycia terenu

Klasteryzacja uwzględnia głównie zależność między wartością oświetlenia terenu o godzinie 9:00 (Hillshade 9am) oraz o godzinie 15:00 (Hillshade 3pm). Najbardziej dominującą grupę stanowi klaster oznaczony kolorem niebieskim, co sugeruje, że większość obserwacji ma zbliżone wartości nasłonecznienia w tych dwóch godzinach. Rozmieszczenie punktów w tym klastrze układają się w charakterystyczny łuk, co może wynikać z naturalnych zależności pomiędzy porannym a popołudniowym oświetleniem terenu. Pozostałe klastry, choć mniej liczne, mogą wskazywać na obszary o specyficznych warunkach terenowych, takich jak różnice w ekspozycji stoków względem słońca lub obecność zacienionych miejsc, gdzie nasłonecznienie jest wyraźnie niższe. Czerwone punkty, będące wartościami odstającymi, mogą reprezentować nietypowe obszary, np. doliny o bardzo ograniczonym dostępie światła słonecznego lub strome stoki o specyficznej orientacji.

5 Klasyfikacja

W celu klasyfikacji kod przygotowuje klastry i porównuje je z klasami faktycznymi – *CoverType*. Dla najlepszych kombinacji zmiennych udało się osiągnąć wyniki w okolicach 35–45%:

- [‘Elevation’, ‘Hillshade_9am’]
- [‘Vertical_Distance_To_Hydrology’, ‘Hillshade_3am’]
- [‘Elevation’, ‘Aspect’, ‘Slope’]

- [‘Vertical_Distance_To_Hydrology’, ‘Horizontal_Distance_To_Hydrology’]

Dla osiągnięcia rezultatów kod zawiera wszystkie dwójkowe i trójkowe kombinacje zmiennych oraz dalsze kombinacje z liczą iteracji (1-15) i wszystkimi trzema metodami klastrów.

W wykorzystaniu kmeans są również 3 metody dostania początkowych centroidów - jedno automatyczne (init) i dwie manualne - losowy punkt i średniony. Oczywiście jest to odpowiednio zrobione dla konkretnych klas (CoverType), dając odpowiednie pokrycie.

Podejście manualne, ręczne było przeprowadzone w celu eksperymentu i porównania wyników, jak i większej zbieżności klastrów i klas faktycznych. Ostatecznie wyniki są delikatnie wyższe dzięki temu, podane powyżej.

```

1  opcja_map = {1: opcja_1, 2: opcja_2, 3: opcja_3}
2  center_func = opcja_map.get(opcja)
3
4  X_sampled = pd.read_csv("dataset.csv", usecols=[a, b, 'Cover_Type'])
5  X_scaled = StandardScaler().fit_transform(X_sampled[[a, b]])
6
7  centers = [
8      center_func(X_sampled[X_sampled['Cover_Type'] == i], a,b)
9      for i in range(1, 8)
10 ]
11 km = KMeans(n_clusters=7, init=centers)
```

Rysunek 22: Kod w edytorze

6 Wnioski i podsumowanie

1. Najlepszą zmienną do badań jest Elevation (Wysokość) i widać tam pewne wzorce, układanie się typów pokrycia w pewnych okolicach choćby:
 - Krummholz na większych wysokościach z mniejszym nachyleniem (można by dodać humorystycznie, że nadrabiają nachylenie kształtem, często mocno wygiętym i nietypowym),
 - Jedlica Douglasa na mniejszych wysokościach, bliżej okolic 2000-2400 metrów,
 - Sosna wydmowa zdecydowanie dominuje w okolicach 3000 metrów i rozciąga na całej długości innych zmiennych (obrazki 10-14).
2. Najmniej wzorców i informacji jest na wykresie 3.1-2 , co pokrywa się z oczekiwaniemi względem tych zmiennych, aczkolwiek mają zaskakującą duże znaczenie w klasyfikacji.
3. Widać, jednak że wykresy cienia wyraźnie układają się w funkcje matematyczne (np. $a^2 - x^2$), co oczywiście ma sens przy poruszaniu się Słońca i kątach promieni o określonych godzinach.
4. Na wykresach ze zmienną Aspect (choćby obrazki 12,13) widać znacznie mniej danych na skrajnych wysokościach i średnim aspekcie, co może sugerować występowanie głębszych geograficznych zależności, którą trzeba by zbadać z większym podłożem wiedzy w tym temacie.

5. Metoda KMeans okazała się najbardziej efektywna i optymalna w obu celach.
6. Metoda DBScan zupełnie nie radzi sobie z tym zbiorem danych i na każdym wykresie widać wyraźną dominację określonego typu, a skuteczność i pokrycie z rzeczywistością są niewielkie.
7. Zaś wykresy metodą Birch mają delikatniejsze przejścia niż Kmeans, mniej liniowe i ograniczone geometrycznie, co zgadza się z oczekiwaniami po specyfice metody, jednak wyniki są odrobinę niższe od wyników KMeans i sama metoda jest mniej optymalna obliczeniowo.

Dane są dość chaotyczne i nie mają bardzo dobrze dobranych klastrów, w celu pełnej analizy czy predykcji warto by się posiąkować innymi metodami, choćby drzewem decyzyjnym i lasami losowymi.

Najlepiej widać to na wykresach 3D, a biorąc to pod uwagę rezultaty są w miarę dobre.

Niestety jest mniej wzorców wśród typów pokrycia i drzew, niż można by oczekiwąć, a sama analiza wyszła dość ograniczona, głównie przez ograniczenie samych metod i wyboru.

Udało się jednak znaleźć parę ciekawych wniosków i informacji, w tym rzeczy warte głębszej analizy.

7 Bibliografia

Literatura

- [1] Blackard, J. A., Dean, D. J. (1999). Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables. *Remote Sensing of Environment*.
- [2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- [3] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.
- [4] Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 226–231.