Raport Projektu EDA

Feature engineering + EDA

Bartłomiej Kózka

8 maja 2025

1 Wstęp

Celem projektu jest wybranie odpowiedniego zestawu danych (nadającego się do wykorzystania w pipeline ML), a następnie wykonanie na nim inżynierii cech oraz eksploracyjnej analizy danych.

Dane

Dane zostały wybrane z serwisu internetowego dane.gov.pl. Ilość danych znajdujących się na tej platformie jest duża, jednak większość tych danych nie nadawała by się do dalszej pracy w pipline ML. Na portalu dane.gov.pl można odnaleźć różne rodzaje danych w wielu kategoriach takich jak: edukacja, energia, gospodarka i finanse, transport i wiele innych. Dane można pobrać w formatach takich jak CSV, JSON, XML, Excel itp. Portal ten oferuje również łatwość w szybkim sprawdzeniu danych przez ich wizualizację (podgląd) na stronie portalu. Dane są również dostępne poprzez REST_API. Dodatkowo można tam również znaleźć krótkie opisy badanych zestawów danych.

Wybór danych

Nabyte dane pochodzą ze stacji meteorologicznej w Brennej z 2020 roku i zawierają zmienne takie jak: ***date*** – data i godzina pomiaru, ***ta*** (temperature) – temperatura powietrza wyrażona w stopniach Celsjusza, ***rh*** (relative humidity) – wilgotność względna w procentach, ***pr*** (pressure) – ciśnienie atmosferyczne w hektopaskalach, ***wv*** (wind velocity) – średnia prędkość wiatru w metrach na sekundę (m/s), ***wpmx*** (wind peak maximum) – maksymalna prędkość wiatru w metrach na sekundę, ***wd*** (wind direction) – kierunek wiatru wyrażony w stopniach oraz ***rf*** (rainfall) – wystąpienie opadów deszczu.

date	ta	rh	pr	wv	wpmx	wd	rf
2020-01-01 00:00	-0,9	94,6	944,6	3,7	7,7	332,2	
2020-01-01 01:00	-0,9	94,3	944,7	3,7	8,3	323,4	0
2020-01-01 02:00	-1	94,2	945	3,2	6,8	323,3	0
2020-01-01 03:00	-1	94	945	3,3	7,2	324,5	0
2020-01-01 04:00	-1,3	94,9	945,2	3,5	7,5	327,8	0
2020-01-01 05:00	-1,5	95,2	945,5	2,9	6	332,6	0
2020-01-01 06:00	-1,7	95,3	945,9	3	5,9	330,2	0

Rysunek 1: Dane meteorologiczne z Brennej

2 Pierwsze Etapy Pipeline'u ML

Wstępna analiza danych polega na zrozumieniu struktury i charakterystyki zbioru danych. W tym kroku przeanalizowano podstawowe informacje o zbiorze, takie jak liczba próbek, liczba cech oraz typy danych. Zbadano również rozkład wartości w poszczególnych kolumnach oraz zidentyfikowano potencjalne problemy, takie jak brakujące wartości czy wartości odstające.

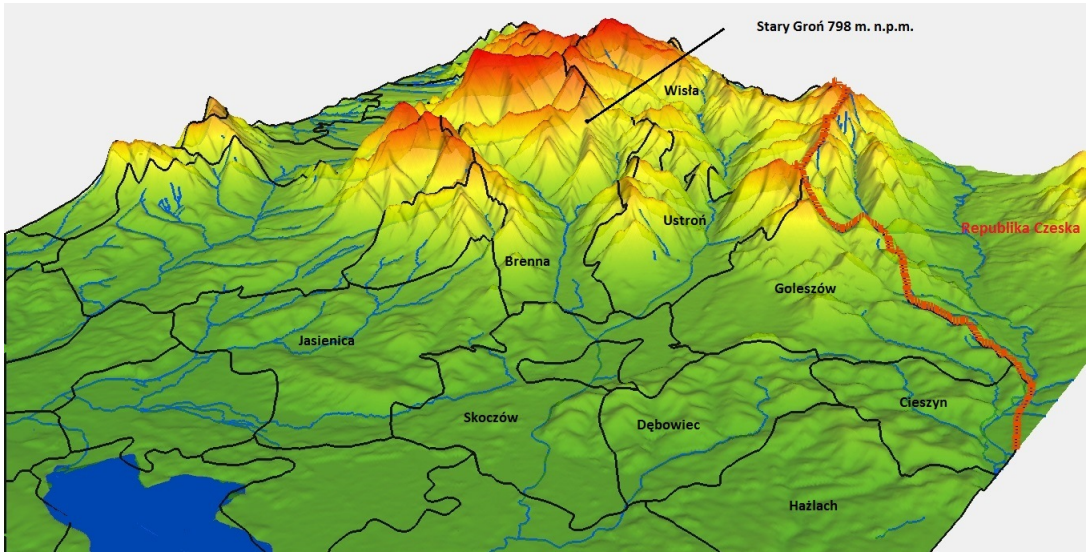
2.1 Inżynieria cech (FE), Eskploracja danych (EDA)

Wstępna analiza oraz cel wykorzystania danych

Dane meteorologiczne mogą być wykorzystane w uczeniu maszynowym do różnych celów, takich jak: prognozowanie warunków pogodowych, klasyfikacja zdarzeń pogodowych, czy też dokładana predykcja wartości ciągłych, np. temperatury lub ilości opadów w mm. W tych przypadkach będziemy mówić o uczeniu nadzorowanym, ponieważ proces ten polega na wykorzystaniu danych wejściowych oraz odpowiednich etykiet, które model ma za zadanie przewidzieć.

Pozyskane dane meteorologiczne z Brennej zawierają 8784 próbek (tyle co liczba godzin w roku przestępnym (366 dni) - 2020 - rok przestępny) oraz 8 cech. Wartości w kolumnach są typu zmiennoprzecinkowego, z wyjątkiem kolumny date, która jest typu datetime64[ns]. Pozostałe zmienne są ciągłe i wyrażają różne parametry meteorologiczne.

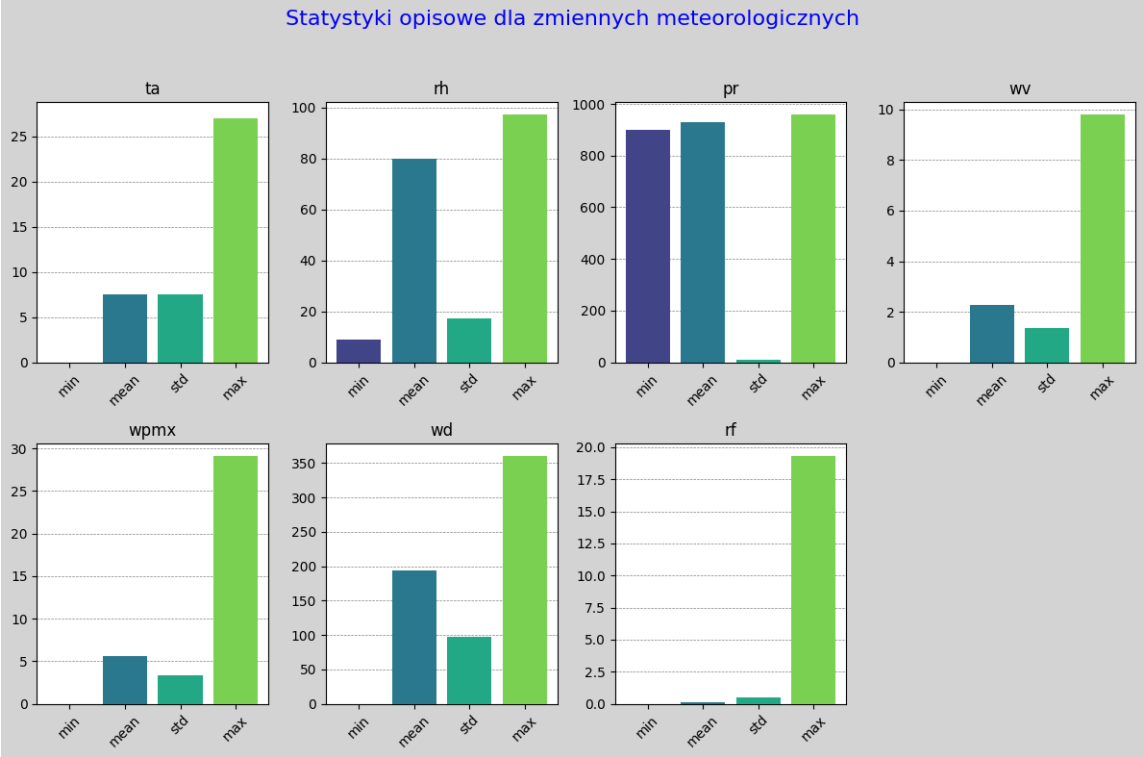
Uwzględniając położenie stacji meteorologicznej w Brennej, która to jest częścią Beskidu śląskiego, z której pochodzą dane, mamy do czynienia z pogodą górską, która to charakteryzuje się duża zmiennością pogody. To znaczy, że warunki atmosferyczne mogą się znacznie różnić od tych w innych regionach Polski. W związku z tym, dane te mogą być szczególnie interesujące dla osób zajmujących się prognozowaniem pogody górskiej, badaniami klimatycznymi czy też dla turystów i mieszkańców regionu.



Rysunek 2: Brenna - lokalizacja

Statystyki opisowe

Dla każdej ze zmiennych obliczono cztery podstawowe miary statystyczne: wartość minimalną, średnią, odchylenie standardowe oraz wartość maksymalną.



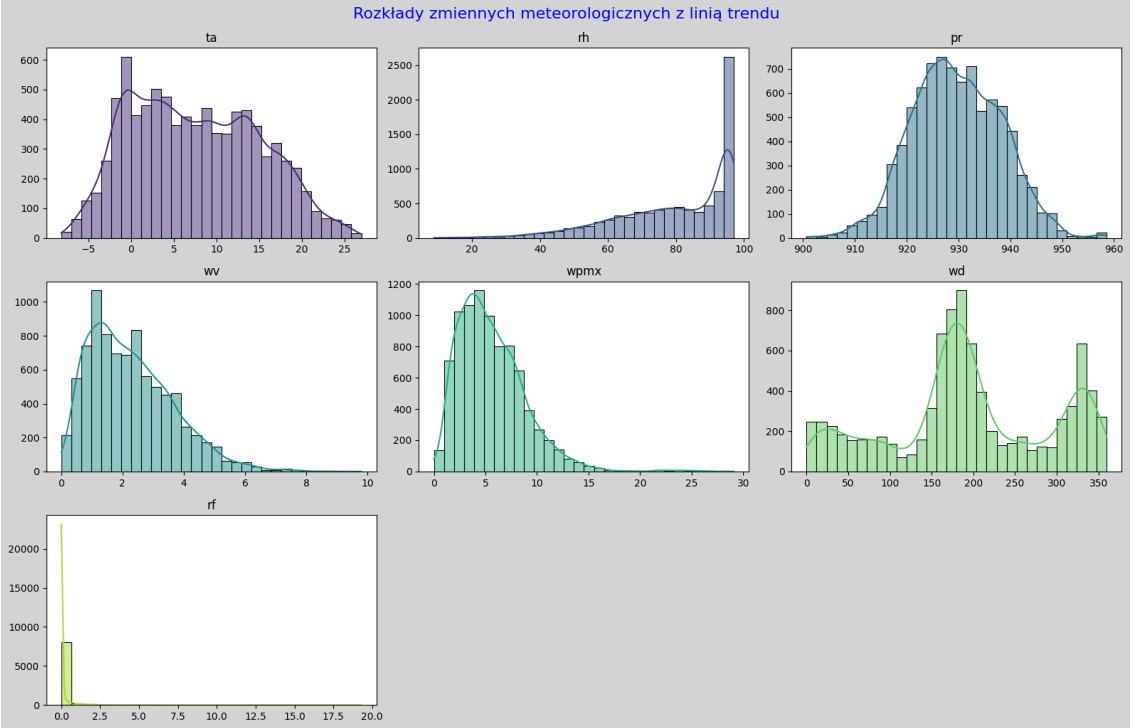
Rysunek 3: Statystyki opisowe dla danych meteorologicznych

Zauważyć można, że temperatura powietrza (ta) osiąga wartości maksymalne powyżej 26°C, natomiast średnia pozostaje w granicach 7°C, co może być wynikiem znacznych wahań temperatury dobowej i sezonowej, typowych dla rejonów górskich. Wilgotność względna (rh) wykazuje dużą zmienność – przy średniej około 80

Prędkość wiatru (wv) i maksymalna prędkość porywu (wpmx), która osiąga blisko 30 m/s to jest ponad 100 km/h sugerują występowanie silniejszych zjawisk wiatrowych, co znajduje uzasadnienie w lokalizacji stacji – doliny i przełęcze sprzyjają przyspieszaniu mas powietrza. Kierunek wiatru (wd) wykazuje dużą rozpiętość wartości, co również jest typowe dla gór, gdzie kierunek wiatru bywa modyfikowany przez ukształtowanie terenu. Zmienna opadów (rf) cechuje się niską średnią, lecz wysoką wartością maksymalną, co potwierdza występowanie intensywnych, epizodycznych opadów – częstych w warunkach górskich, szczególnie w okresie letnim.

Podsumowując, przedstawione dane dostarczają cennych informacji na temat lokalnych warunków klimatycznych w Beskidzie Śląskim. Ich analiza jak już wcześniej zostało wspomniane może służyć jako podstawa do dalszych badań nad mikroklimatem górskim.

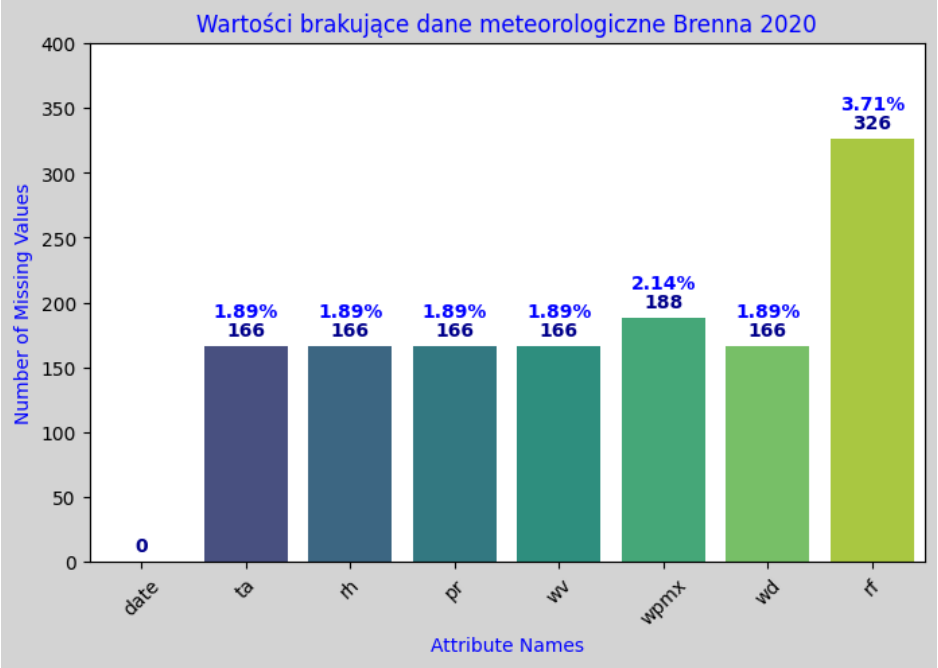
Rozkłady cech



Rysunek 4: Rozkład cech danych meteorologicznych

Spośród wszystkich analizowanych zmiennych szczególnie ciekawy rozkład wykazuje zmienna wd, odpowiadająca za kierunek wiatru. Zamiast jednolitego lub przypadkowego rozkładu, widoczne są wyraźne piki w kilku konkretnych kierunkach, co sugeruje, że w Brennej wiatry najczęściej wieją z określonych stron świata. Taki układ może odzwierciedlać lokalne uwarunkowania geograficzne, np. obecność dolin, grzbietów górskich czy wpływ dominujących mas powietrza, tak jak powyżej było to już wspomniane.

Wartości brakujące



Rysunek 5: Wartości brakujące w danych meteorologicznych

Jak można zaobserwować na powyższym histogramie, tylko w kolumnie oznaczającej datę i godzinę pomiaru (date) nie występują żadne wartości brakujące co jest zgodne z oczekiwaniami, ponieważ jest to zmienna czasowa, która powinna być kompletna. Pozostałe zmienne zawierają pewną liczbę brakujących wartości. W przypadku zmiennych takich jak temperatura (ta), wilgotność (rh), ciśnienie (pr), prędkość wiatru (wv), maksymalna prędkość wiatru (wpmx), kierunek wiatru (wd), liczba brakujących danych wynosi 166, co stanowi niewielki procent w porównaniu do całkowitej liczby próbek. Dla zmiennej rf (opady) brakujących wartości jest 326, co również jest małym procentem w odniesieniu do całkowitej liczby danych. W związku z tym, brakujące dane są stosunkowo niewielkie i nie mają istotnego wpływu na całościową strukturę danych.

Biorąc pod uwagę niewielki procent brakujących danych oraz fakt, że braki są rozłożone równomiernie, można przyjąć, że brakujące wartości w danych meteorologicznych mają charakter **Missing Completely at Random (MCAR)**. Oznacza to, że brakujące dane są losowe i nie zależą od innych zmiennych w zbiorze danych. Prawdopodobnie przyczyną braków odczytów były zakłócenia w pracy czujników lub inne techniczne problemy związane z urządzeniami pomiarowymi.

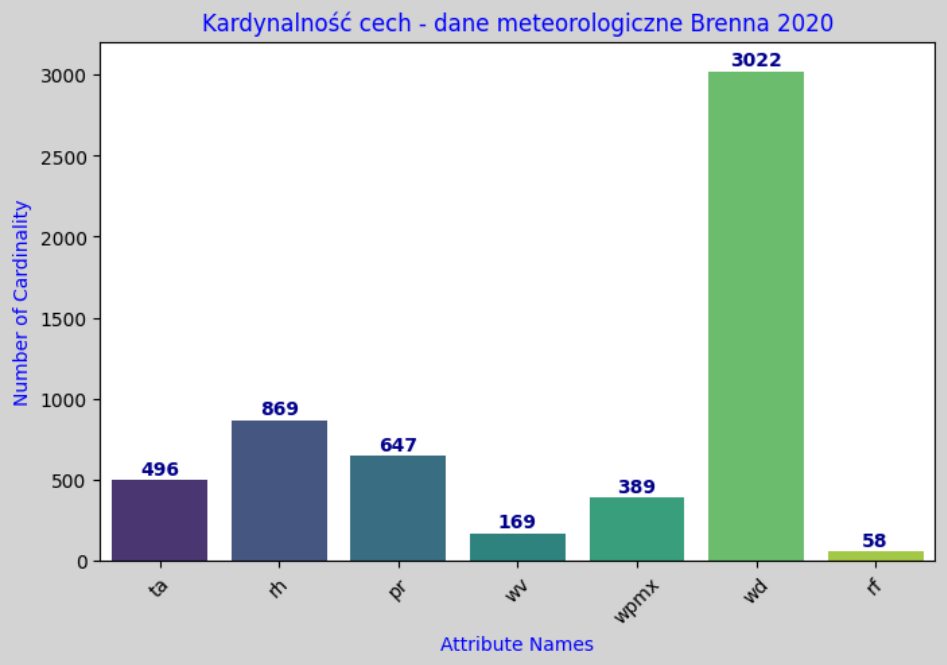
W celu przygotowania danych meteorologicznych do modelowania zastosowano dwie różne metody uzupełniania brakujących wartości, odpowiednio do charakteru poszczególnych zmiennych. Dla zmiennych ciągłych takich jak temperatura powietrza (ta), wilgotność względna (rh), ciśnienie atmosferyczne (pr), prędkość wiatru (wv), maksymalna prędkość porywu wiatru (wpmx) oraz kierunek wiatru (wd) zdecydowano się na interpolację liniową w funkcji czasu. Metoda ta pozwala na płynne oszacowanie brakujących wartości na podstawie danych sąsiednich, co odzwierciedla naturalne zmiany tych parametrów w czasie i zapewnia spójność czasową pomiarów.

W przypadku zmiennej dotyczącej opadów (rf) przyjęto inne podejście: brakujące wartości zostały zastąpione zerami, ponieważ brak danych może oznaczać brak

opadów, co jest logicznie uzasadnione i nie wprowadza istotnych zniekształceń do analizy.

Kardynalność cech

Została obliczona kardynalność cech, czyli liczba unikalnych wartości dla każdej z cech. Na poniższym wykresie celowo kardynalość cechy date została pominięta, ponieważ jest to zmienna czasowa, która z definicji ma unikalne wartości dla każdej godziny w roku. Pozostałe zmienne mają różną kardynalność, co może wpływać na dalszą analizę i modelowanie. Wartości kardynalności dla poszczególnych zmiennych przedstawione są na poniższym wykresie.



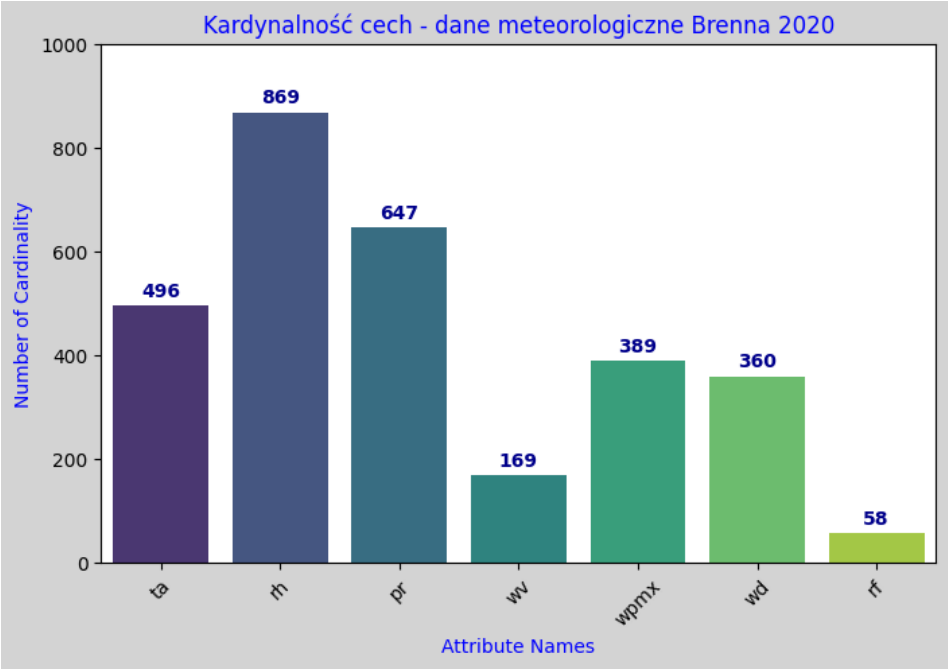
Rysunek 6: Kardynalność cech przed redukcją

W przedstawionych wynikach liczby unikalnych etykiet zmiennych, zauważalna jest szczególnie wysoka kardynalność w przypadku zmiennej wd (kierunek wiatru), która liczy aż 3022 unikalnych wartości. Taki wynik jest wynikiem precyzyjnego pomiaru w stopniach, co skutkuje dużą liczbą unikalnych kategorii. Pozostałe zmienne, takie jak temperatura (ta), wilgotność (rh), ciśnienie (pr), prędkość wiatru (wv), maksymalna prędkość porywu wiatru (wpmx) oraz opady (rf), wykazują raczej umiarkowaną liczbę unikalnych wartości, co jest typowe dla danych meteorologicznych i nie stanowi problemu w dalszej analizie.

W celu uproszczenia analizy danych meteorologicznych, zdecydowano się na redukcję zmiennej wd (kierunek wiatru). Z uwagi na fakt, iż jest to cecha, która nie wyamaga precyzji co do dziesiętnej części ułamka do predykcji jej samej (co i tak może być bradzo cieżkim wyzwaniem), jak i zarówno w kontekście bycia zmienną objaśniającą dla innnej zmiennej objaśnianej, postanowiono zredukować jej precyzję na liczby całkowitę co będzie w zupełności wystarczalne w dalszej częsci pipline'u ML. Można by również pomyśleć nad redukcją tej zmiennej do 8 kategorii (N, NE, E, SE, S, SW, W, NW), jednakże w tym przypadku nie jest to konieczne.

Pozostałe zmienne, takie jak temperatura, wilgotność czy ciśnienie, pozostają w oryginalnej formie, ponieważ zawierają cenne informacje, które mogą wpływać na przewidywania związane z opadami i innymi zjawiskami meteorologicznymi.

Wartościową operacją w kontekście dalszego przebiegu pipeline'u mogło by również być zredukowanie cechy rf (rainfall) na zmienną binarną (0 lub 1), co oznaczało by brak opadów lub ich występowanie. Jednakże na tym etapie pipeline'u nie jest to wskazane, ze względów na brak wiedzy co do dalszego wykorzystania danych. W przypadku chęci predykcji jak duże są opady w zależnoći od innych zmiennych, sprowadzenie tej zmiennej do wartości binarnej było by kategorycznym błędem.



Rysunek 7: Kardynalność cech po redukcji zmiennej wd

FE - Dodanie wartościowych zmiennych

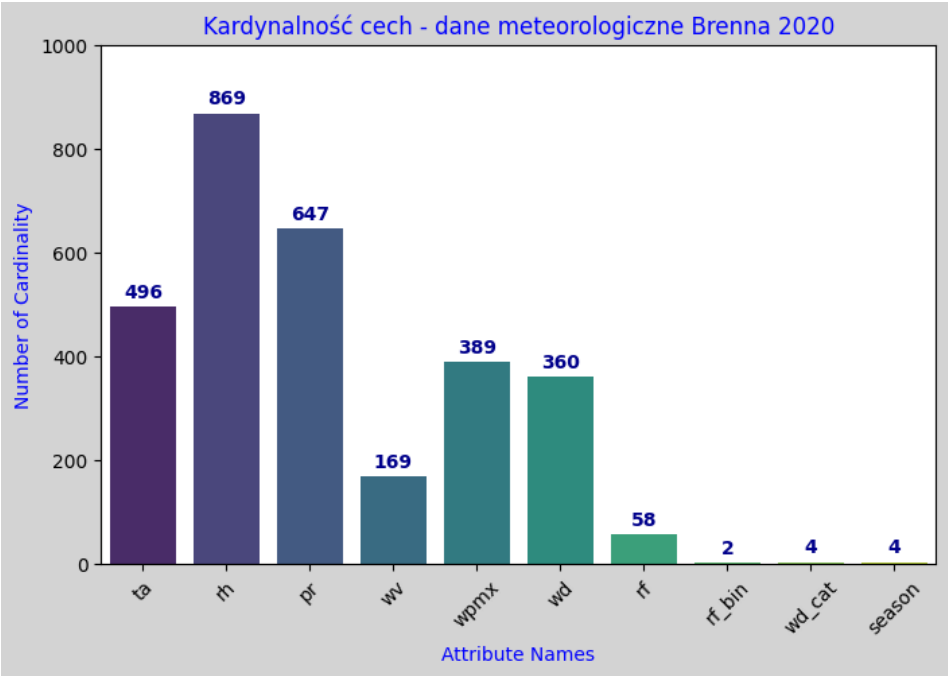
Po przeanalizowaniu oraz redukcji kardynalości cech, zdecydowano się na dodanie nowych zmiennych, które mogą być istotne w kontekście dalszej analizy i modelowania. Wprowadzono następujące zmienne:

season (pora roku) – zmienna kategoryczna, która wskazuje na porę roku (wiosna, lato, jesień, zima). Zmienna ta została dodana w celu uwzględnienia sezonowości w danych meteorologicznych. W różnych porach roku warunki atmosferyczne mogą się znacznie różnić, co może mieć wpływ na występowanie opadów i inne zjawiska pogodowe.

wd_cat (kierunek wiatru ze względu na kierunki geograficzne) – zmienna kategoryczna, która wskazuje na kierunek wiatru w czterech głównych kierunkach geograficznych (N, E, S, W). Zmienna ta została dodana w celu uproszczenia analizy kierunku wiatru i umożliwienia lepszego zrozumienia jego wpływu na inne zmienne meteorologiczne, bez konieczności wchodzenia w szczegółowy (kątowy) opis kierunku waitru.

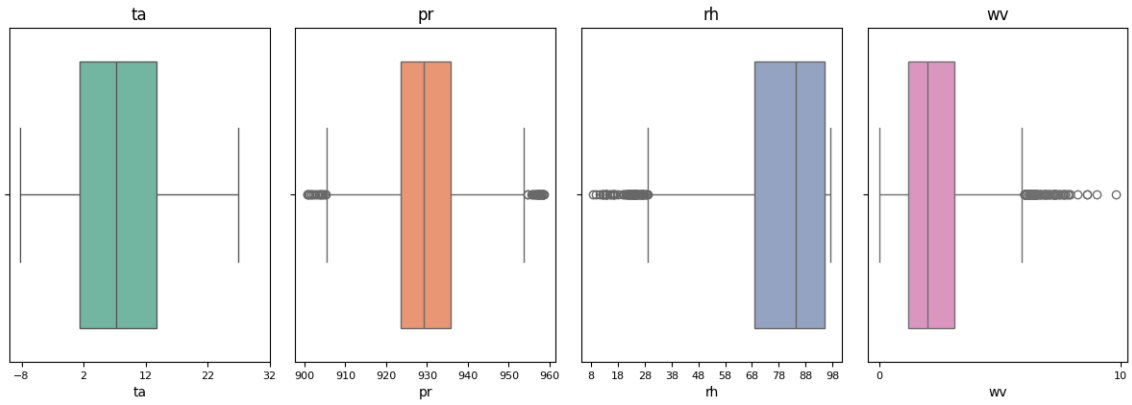
rf_bin opady deszczu (0 lub 1) – zmienna binarna, która wskazuje na wystąpienie opadów deszczu (1) lub ich brak (0). Zmienna ta została również doda-

na w celu uproszczenia analizy i umożliwienia lepszego zrozumienia wpływu innych zmiennych na występowanie opadów, np. w kontekśście przewidywania czy opady miały miejsce czy nie.



Rysunek 8: Kardynalność cech po dodaniu nowych zmiennych

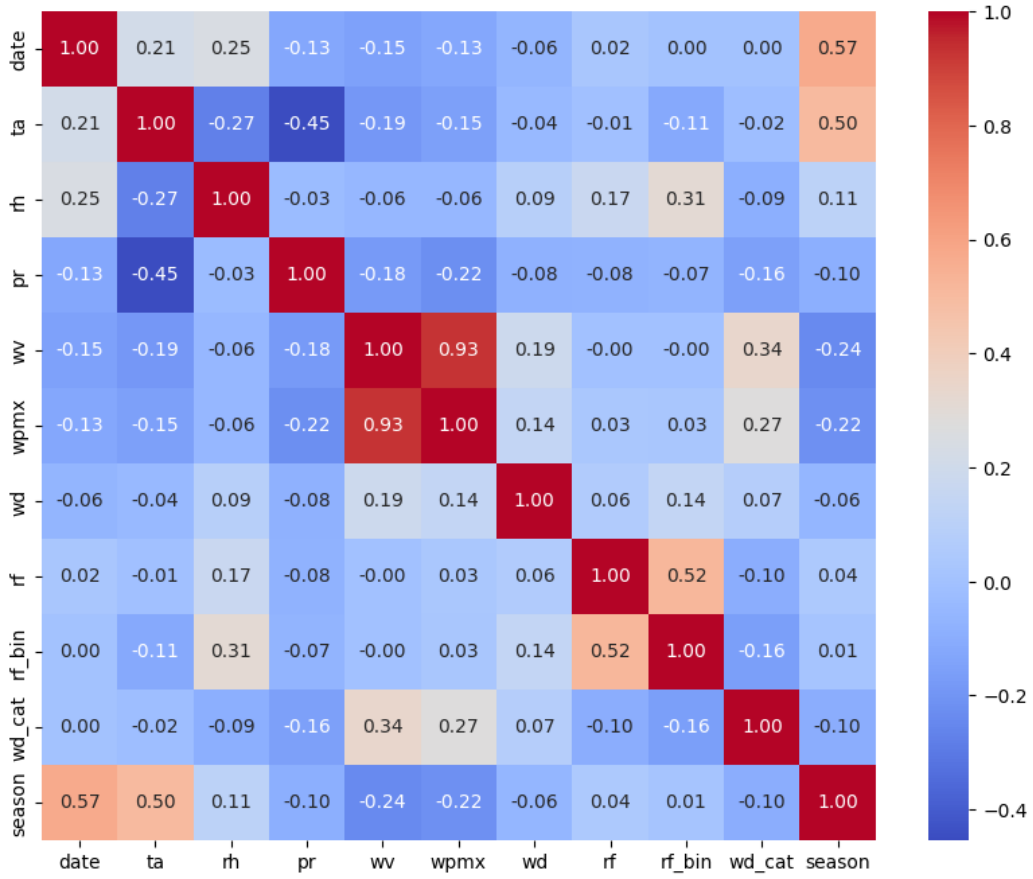
Wartości odstające



Rysunek 9: Wartości odstające

Po przeprowadzeniu analizy danych w celu odnalezniie anomalii nie wykryto obecności wyraźnych wartości odstających (outlierów) w zestawie danych meteorologicznych. Wszystkie zmienne rozkładają się w sposób typowy dla danych tego typu, a wartości skrajne, jeśli występują, nie odbiegają znacząco od ogólnych wzorców. W związku z tym, dane zostały uznane za stabilne i nie wymagały dalszej obróbki w zakresie usuwania wartości odstających.

Macierz korelacji



Rysunek 10: Korelacja pomiędzy zmiennymi

Na podstawie analizy można zauważyć silną dodatnią korelację między średnią prędkością wiatru (wv) a jego maksymalną wartością (wpmx), osiągającą wartość 0,93. Wskazuje to na dużą współzmienność obu zmiennych – dni z wyższą średnią prędkością wiatru charakteryzują się również wyższym wiatrem maksymalnym. Umiarkowanie silna ujemna korelacja (r = -0,45) występuje między temperaturą powietrza (ta) a ciśnieniem atmosferycznym (pr), co może odzwierciedlać wpływ lokalnych warunków pogodowych oraz sezonowych układów ciśnienia. Zmienna sezonowa (season), wyodrębniona z daty, wykazuje dodatnią korelację zarówno z temperaturą (r = 0,50), jak i z samą datą (r = 0,57), co potwierdza wpływ pory roku na rozkład temperatury i chronologiczną strukturę danych. Zmienna opadowa (rf) oraz jej wersja binarna (rf_bin) również wykazują ze sobą silną zależność (r = 0,52), co jest spodziewane ze względu na logiczną konwersję wartości ciągłej na kategorię obecności lub braku opadu. Pozostałe korelacje między zmiennymi są słabe lub bardzo słabe, co sugeruje niewielkie liniowe zależności bądź możliwość występowania relacji nieliniowych, niewidocznych w analizie korelacji Pearsona.

2.2 Uczenie nadzorowane

Wybór zmiennej Target

Jako zmienną TARGET wybrano **temperaturę powietrza (ta)**. Temperatura jest jedną z kluczowych wielkości meteorologicznych — wpływa na życie ludzi, roślinność i cykle przyrody. Jest też najczęściej prognozowaną zmienną w meteorologii. W warunkach klimatu górskiego, gdzie występują duże dobowe i sezonowe wahania temperatury, jej przewidywanie jest szczególnie istotne, m.in. z powodu wpływu na lokalne zjawiska pogodowe, transport czy rolnictwo. Ukształtowanie terenu (np. doliny, stoki) dodatkowo wzmacnia lokalne różnice temperatur, co sprawia, że jest to naturalny wybór jako zmienna do modelowania.

Wybór zmiennych Features

Wybrane zmienne FEATURES to:
Ciśnienie (pr) – zmiany ciśnienia wiążą się z przemieszczaniem mas powietrza i frontów atmosferycznych, które w rejonach górskich silnie wpływają na zmiany temperatury.
Wilgotność (rh) – powietrze w górach często ma zmienną wilgotność; duża wilgotność może ograniczać nagrzewanie, a suchy klimat sprzyja szybkiemu nagrzewaniu się powietrza.
Wiatr (wv, wpmx) – w terenie górskim wiatr pełni ważną rolę w przenoszeniu mas powietrza między wysokościami i dolinami, co bezpośrednio wpływa na temperaturę.
Pora roku (season) – sezon wpływa na długość dnia, nasłonecznienie oraz częstość występowania zjawisk takich jak inwersje temperatury, które są typowe dla dolin górskich.

Nie uwzględniono kierunku wiatru (wd) ani opadów (rf), ponieważ ich wpływ na temperaturę jest bardziej pośredni i trudniejszy do jednoznacznego ujęcia.

Źródła

- `https://brenna.meteo.com.pl/`