Politechnika Wrocławska

Predykcja ekstremalnych zjawisk meteorologicznych

Metody analizy i eksploracji danych

Wykonali: Przemysław Marciniak 247331

Jakub Mazur 247379

Wydział: *Informatyki i Telekomunikacji W-4n*

Rok Akademicki : *2022/2023*

Grupa: *K01-36a*
Termin: *czwartek 17:05*

Prowadzący: ***Dr inż. Agata Migalska***

Baza danych

Baza danych wykorzystana w ramach realizacji projektu została pozyskana z publicznego archiwum Instytutu Meteorologii i Gospodarki Wodnej (https://danepubliczne.imgw.pl/data/). Zawarte w niej są szczegółowe obserwacje składników pogody na terenie całej Polski (wstępnie wykorzystane zostały dane pochodzące z terenów Warszawy). Pomiary wykonywane były codziennie od 1960 r. do 2022 r. W bazie znajdują się następujące atrybuty:

- kod stacji
- nazwa stacji
- rok
- miesiąc
- dzień
- max temp. d. [°C]
- min temp. d. [°C]
- śr. temp. d. [°C]
- temp. min przy gruncie [°C]
- suma d. opadu [mm]
- rodzaj opadu [S/W/]
- wysokość pokrywy śnieżnej [cm]
- równoważnik wodny śniegu [mm/cm]
- czas usłonecznienia [godz.]
- czas opadu deszczu [godz.]
- czas opadu śniegu [godz.]
- czas opadu deszczu ze śniegiem [godz.]
- czas gradu [godz.]
- czas mgły [godz.]
- czas zamglenia [godz.]
- czas sadzi [godz.]
- czas gołoledzi [godz.]
- czas zamieci śnieżnej niskiej [godz.]
- czas zamieci śnieżnej wysokiej [godz.]
- czas zmętnienia [godz.]
- czas wiatru >= 10 m/s [godz.]
- czas wiatru > 15 m/s [godz.]
- czas burzy [godz.]
- czas rosy [godz.]
- czas szronu [godz.]
- wystąpienie pokrywy śnieżnej [0/1]
- wystąpienie błyskawicy [0/1]
- stan gruntu [Z/R]
- izoterma dolna [cm]
- izoterma górna [cm]
- aktynometria [J/cm^2]
- śr. d. zachmurzenie ogólne [oktanty]
- śr. d. prędkość wiatru [m/s]
- śr. d. ciśnienie pary wodnej [hPa]
- śr. d. wilgotność względna [%]
- śr. d. ciśnienie na poziomie stacji [hPa]
- śr. d. ciśnienie na poziomie morza [hPa]
- suma opadu w dzień [mm]
- suma opadu w noc [mm]

Dodatkowo do każdego z pomiarów przypisany jest atrybut „status pomiaru" przyjmujący wartości [8/9/], gdzie 8 oznacza brak pomiaru, 9 brak wystąpienia zjawiska, a w przypadku pustego pola, wystąpienie zjawiska i dokonanie pomiaru.

Dane z każdego roku dostarczone były w dwóch różnych plikach .csv zawierających informacje o różnych składnikach pogody. Z tego względu na wstępie potrzebne było ujednolicenie plików do jednej bazy danych, a następnie usunięcie zduplikowanych kolumn. Wykluczone zostały wyniki z lat 1960-1965, ze względu na całkowity brak pomiarów wszystkich parametrów (status pomiaru równy 8), a także atrybuty wnoszące znikomą ilość informacji w kwestii badanych zjawisk meteorologicznych. Z pośród wyników jedynie znikoma liczba wzbudziła podejrzenie co do swojej poprawności. Pomiary, które w zdecydowany sposób odbiegały od normy, znalazły potwierdzenie w archiwalnych artykułach lub prognozach pogody. Na przykład, w przypadku największego, odstającego, pomiaru wysokości pokrywy śnieżnej, udało znaleźć się artykuł (https://warszawa.naszemiasto.pl/zima-stulecia-w-warszawie-tak-bylo-40-lat-temu-zobaczcie/ar/c8-4356201) opisujący wystąpienie tego zdarzenia. Ostatecznie otrzymana baza danych składników pogody dla Warszawy posiada 48 atrybutów dla 20454 rekordów.

Eksploracyjna analiza danych

Potencjalnie niebezpieczne zjawiska pogodowe, których predykcja podjęta została w ramach projektu to: ulewne deszcze, śnieżyce, wichury, burze oraz ekstremalne temperatury (mróz i upał). Granice wartości od których odczytane wyniki uznawane zostają za anomalię przyjęto na podstawie meteorologicznych stopni zagrożenia. W trakcie analizy zjawisk została wygenerowana duża ilość wykresów pokazujących zależności między zmiennymi, w celu zachowania czytelności raportu zamieszczona została tylko część z nich.

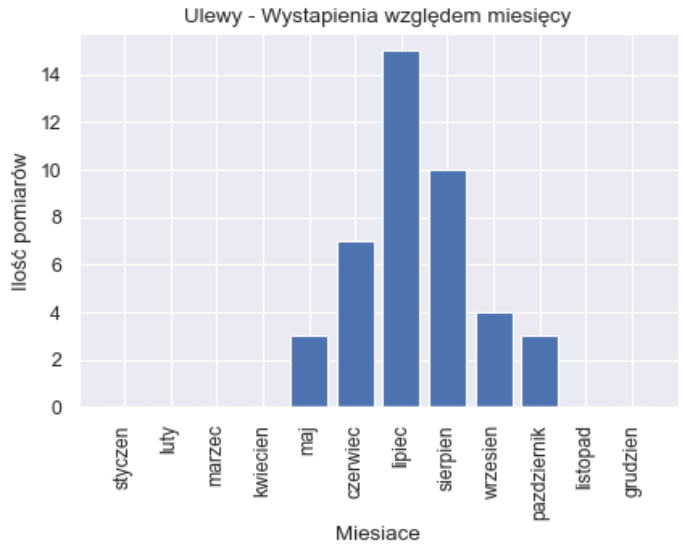
Ulewa - zjawisko charakteryzujące się dużymi opadami dobowymi deszczu, przekraczającymi 30mm. Na potrzeby analizy tego zjawiska, utworzony został podzbiór który zawiera wyłącznie dni kiedy: a) Pomiar został wykonany poprawnie, b) Opad miał charakter deszczu, c) Suma dobowa opadu wyniosła ponad 30mm.

Wstępnie badane zależności między zmiennymi:

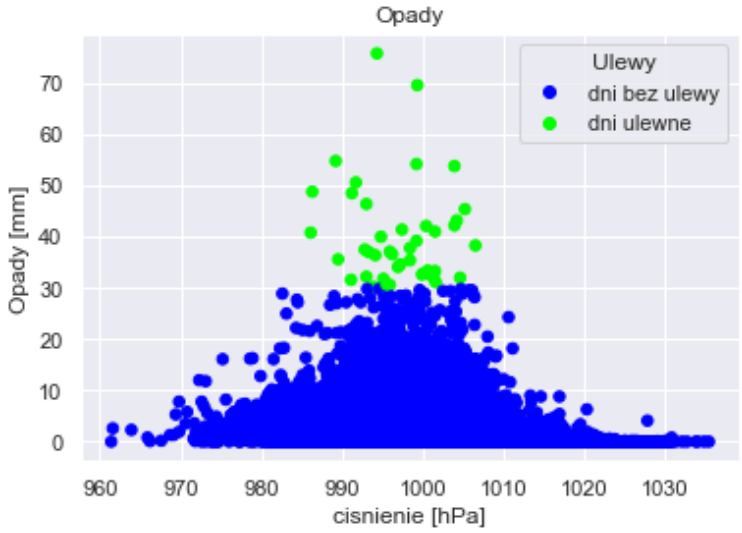
1. Rozkład ilości wystąpień zjawiska w zależności od roku oraz od miesiąca,
2. Suma dobowa opadów od ciśnienia atmosferycznego na poziomie stacji,
3. Średnie zachmurzenie od średniej dobowej wilgotności.

Przykładowe zależności:

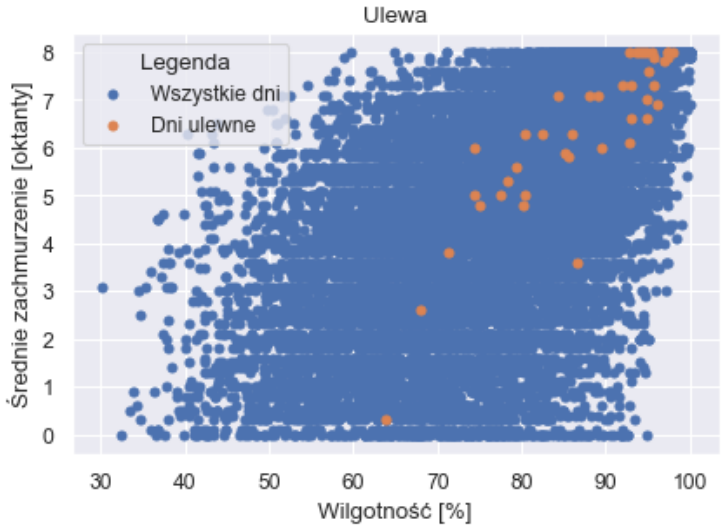
Rysunek 1 Liczba wystąpień ulew w kolejnych latach



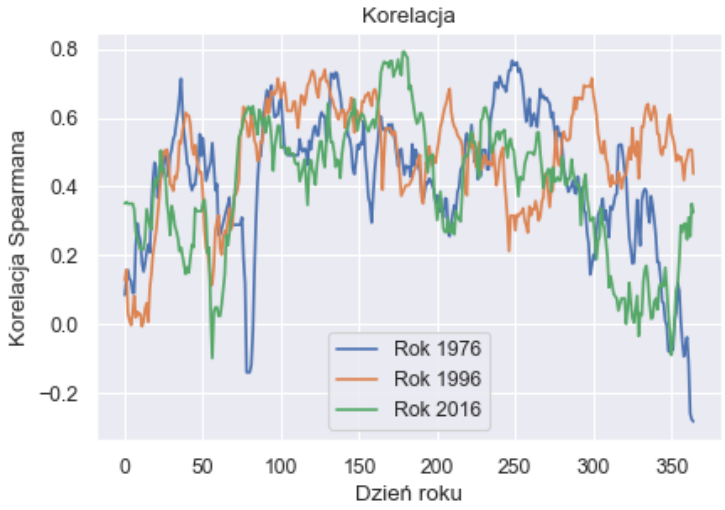
Rysunek 2 Liczba wystąpień ulew według miesięcy



Rysunek 3 Rozkład opadów w zależności od ciśnienia



Rysunek 4 Rozkład ulew w zależności od średniej dobowej wilgotności powietrza i średniego zachmurzenia



Rysunek 5 Korelacja Spearmana wykonana ruchomym, 30 dniowym oknem, dla 3 lat (1976,1996,2016)

Korelacje Pearsona i Spearmana prezentują się następująco:

Tabela 1 Współczynniki Pearsona i Spearmana dla ulewy

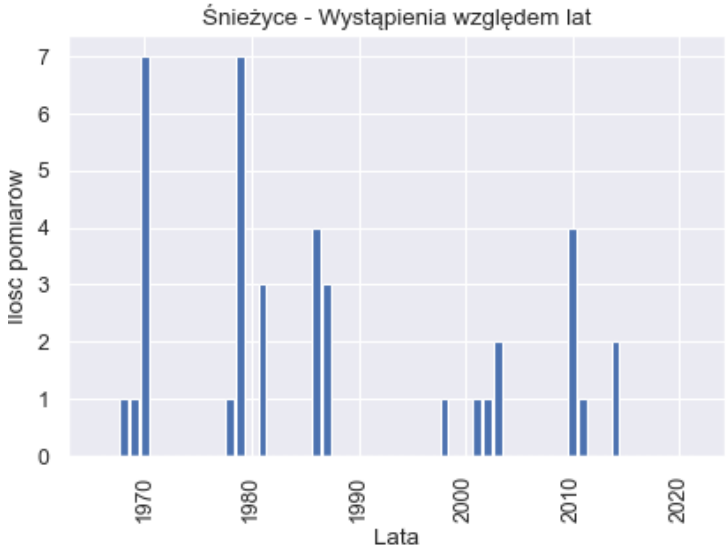
Korelacja wysokości opadów z:	Współczynnik Pearsona	Współczynnik Spearmana
Ulewy – Dzienna różnica ciśnień	0.1617	0.2536
Ulewy - Ciśnienie	-0.0924	-0.0528
Ulewy – Średnia temp.	0.1403	0.1770
Ulewy - Wilgotność	0.0953	-0.0704
Cały zbiór – Dzienna różnica ciśnień	-0.1498	-0.2023
Cały zbiór – Ciśnienie	-0.2412	-0.4259
Cały zbiór – Średnia temp.	0.1178	-0.0292
Cały zbiór – Wilgotność	0.2062	0.3774

Śnieżyca - zjawisko analogiczne do ulewy, z tym wyjątkiem, że zamiast opadów deszczu występują opady śniegu. Sytuację anormalną przyjmujemy w momencie gdy: a) wysokość pokrywy śnieżnej utrzymuje się na poziomie większym niż 15 cm, b) czas trwania wiatru o prędkości co najmniej 10m/s przekracza 1 godzinę.

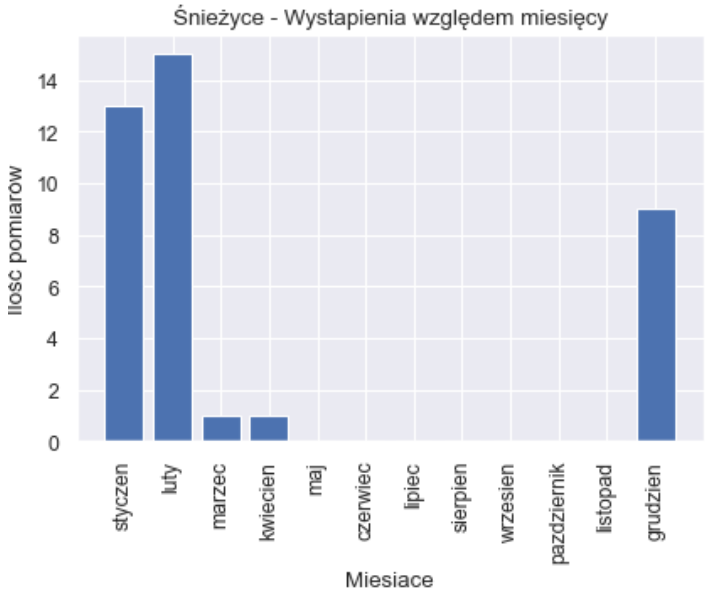
Wstępnie badane zależności między zmiennymi:

1. Rozkład ilości wystąpień zjawiska w zależności od roku oraz od miesiąca,
2. Wysokość pokrywy śnieżnej od średniej temperatury dobowej,

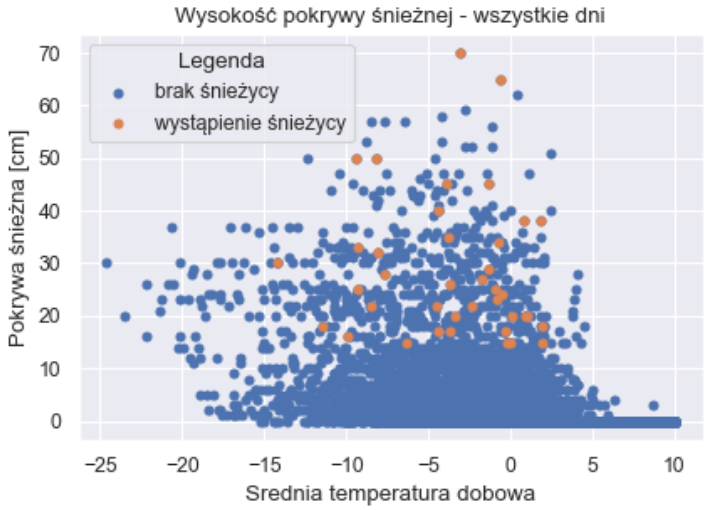
Przykładowe wykresy i histogramy:



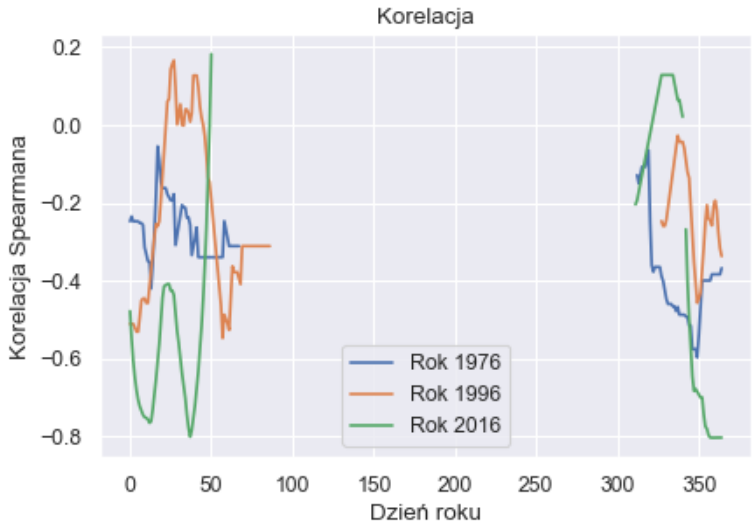
Rysunek 6 Liczba wystąpień śnieżyc w kolejnych latach



Rysunek 7 Liczba wystąpień śnieżyc według miesięcy



Rysunek 8 Wykres zależności wysokości pokrywy śnieżnej od średniej temperatury dobowej



Rysunek 9 Wykres korelacji wysokości pokrywy śnieżnej w zależności od temperatury (Spearman – 30dniowe okno)

Korelacje Pearsona i Spearmana prezentują się następująco:

Tabela 2 Współczynników Pearsona i Spearmana dla śnieżycy

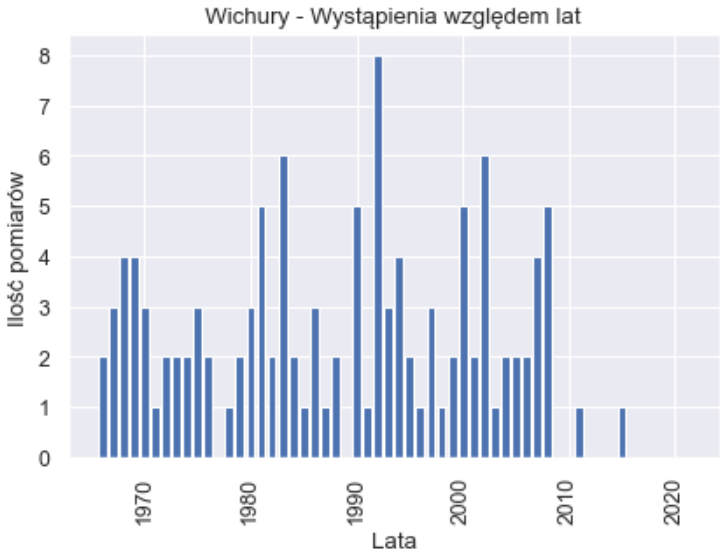
Korelacja wysokości opadów z:	Współczynnik Pearsona	Współczynnik Spearmana
Temp - Wysokość pokrywy śnieżnej - śnieżyce	-0.0825	-0.19744
Wilgotność - Wysokość pokrywy śnieżnej - śnieżyce	-0.2954	-0.3575
Temp - Wysokość pokrywy śnieżnej – Cały zbiór	0.1178	-0.0292
Wilgotność - wysokość pokrywy śnieżnej – Cały zbiór	0.2062	0.3774

Na podstawie rysunku 9, zauważyć można, że w momencie gdy w przedziale analizowanym oknem nie znajduje się żadne wystąpienie zjawiska, wykres ulega zerwaniu. Dla zjawisk pogodowych charakterystycznych dla danej pory roku (np. śnieżyca), przerwanie wykresu na okres wiosna-jesień jest logiczne, jednak w momencie gdy przebieg nie jest ciągły w okresie występowania zjawiska (patrz rysunek 9 Rok 2016) powoływanie się na wykres korelacji traci wiarygodność i sens. Sytuacja powtarza się dla każdego innego zjawiska pogodowego, chyba że wybrane zostanie odpowiednio duże okno czasowe (sięgające nawet 30 dni jak na rysunku 5). W związku z tym w dalszej części raportu zaniechana została analiza przebiegu korelacji. Ponadto ze względu na dużą zmienność uzyskanych wartości (tabela 1 i 2) w zależności od analizowanego zbioru (brak konsekwencji), zaniechane zostało wyznaczanie współczynników korelacji.

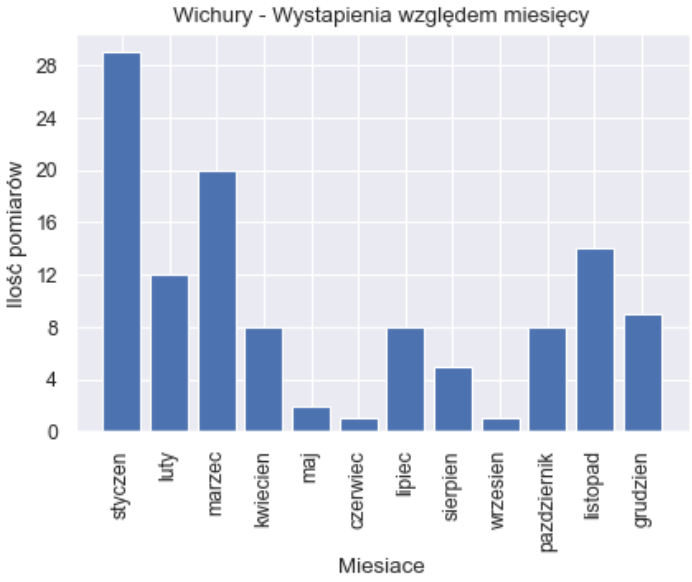
Wichura - silny wiatr o prędkości powyżej 15 m/s. Sytuację anormalną przyjmujemy gdy czas trwania wiatru >15 m/s był większy niż 0 h. W tym przypadku badane są zależności pomiędzy:

1. Rozkładem ilości wystąpień zjawiska w zależności od roku oraz od miesiąca,
2. Zmianą średniego ciśnienia pomiędzy dwoma kolejnymi dniami i średnią temperaturą dobową.

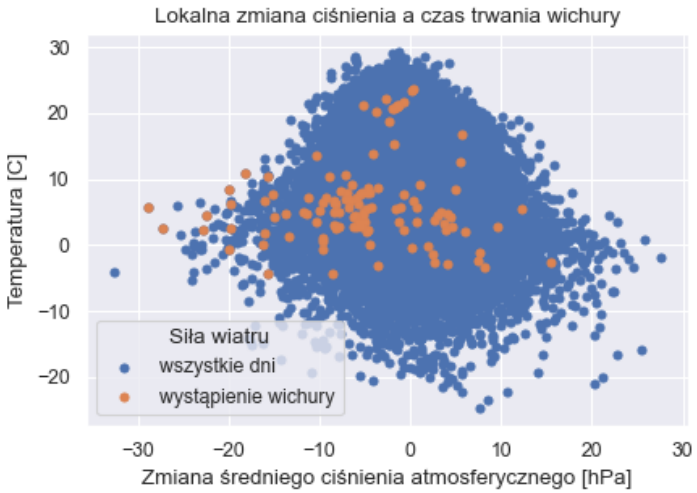
Przykładowy wykres i histogram:



Rysunek 10 Liczba wystąpień wichur w kolejnych latach



Rysunek 11 Liczba wystąpień wichur według miesięcy

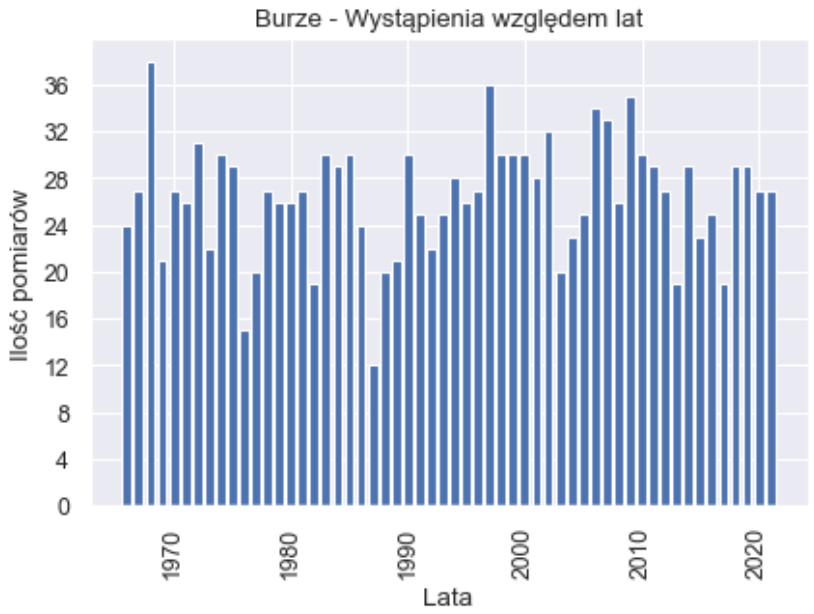


Rysunek 12 Wykres zależności wystąpienia wichury, a zmiany ciśnienia atmosferycznego i średniej temperatury.

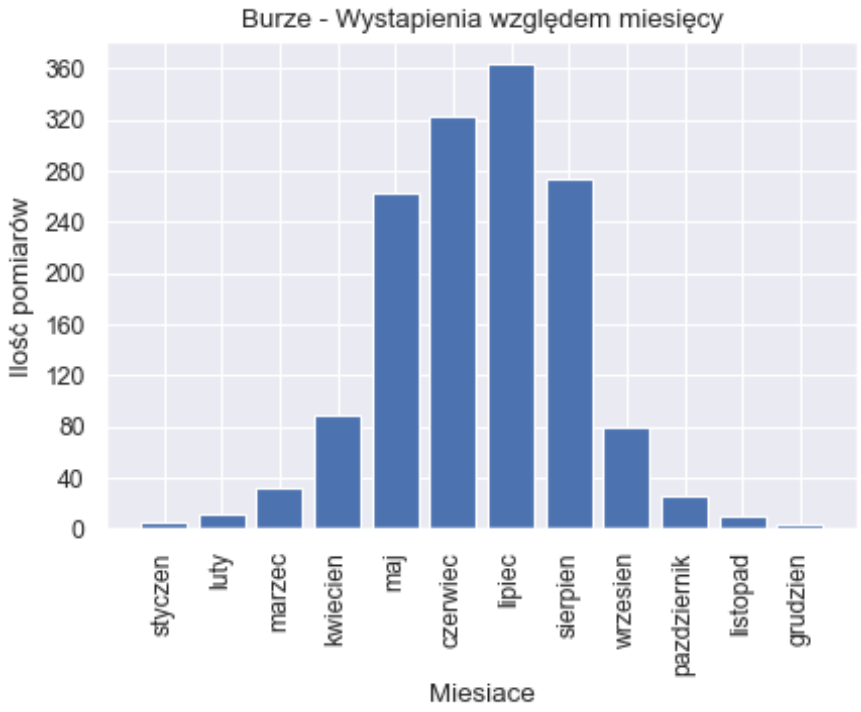
Burza - charakteryzuje się występowaniem dużych opadów deszczu oraz silnymi wiatrami, opcjonalnie mogą pojawić się błyskawice. Sytuację anormalną przyjmujemy gdy czas trwania burzy był większy niż 0 h. Rozważane są zależności pomiędzy:

1. Rozkład ilości wystąpień zjawiska w zależności od roku oraz od miesiąca,
2. Średnią wilgotnością względną a średnią temperaturą dobową.

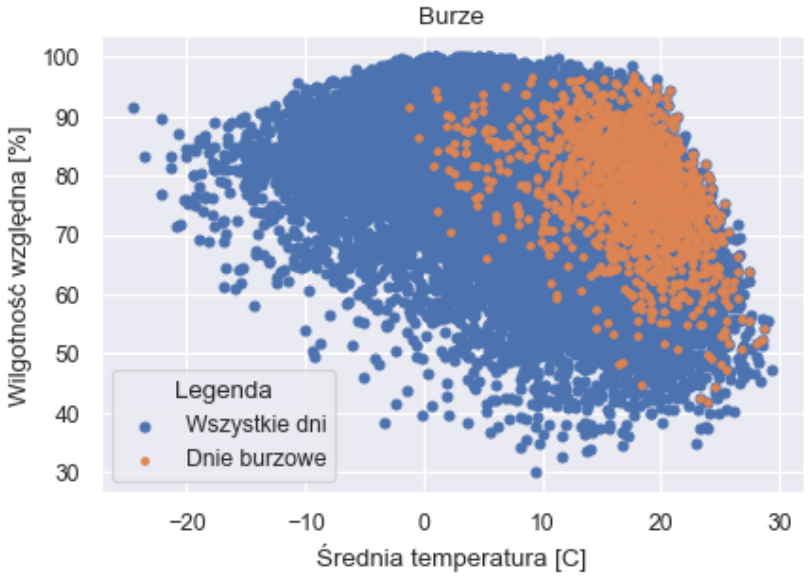
Przykładowe wykresy i histogramy:



Rysunek 13 Liczba wystąpień burz w kolejnych latach



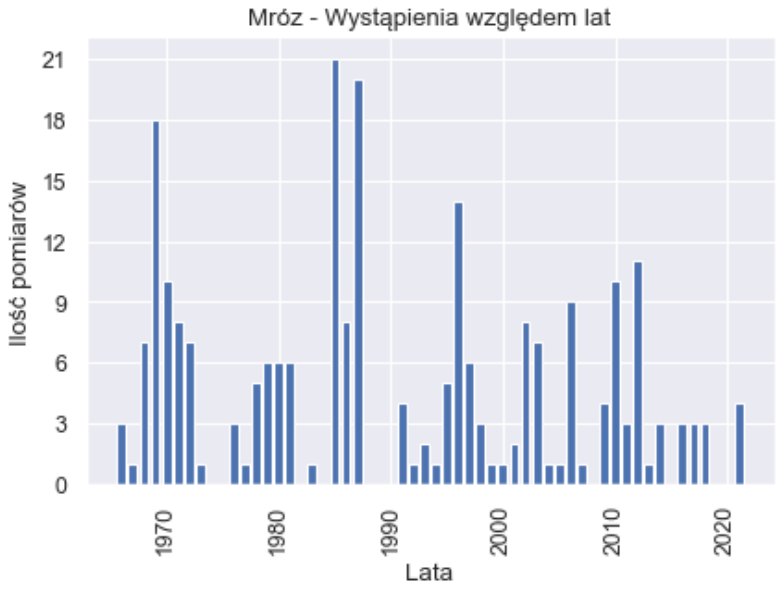
Rysunek 14 Liczba wystąpień burz według miesięcy



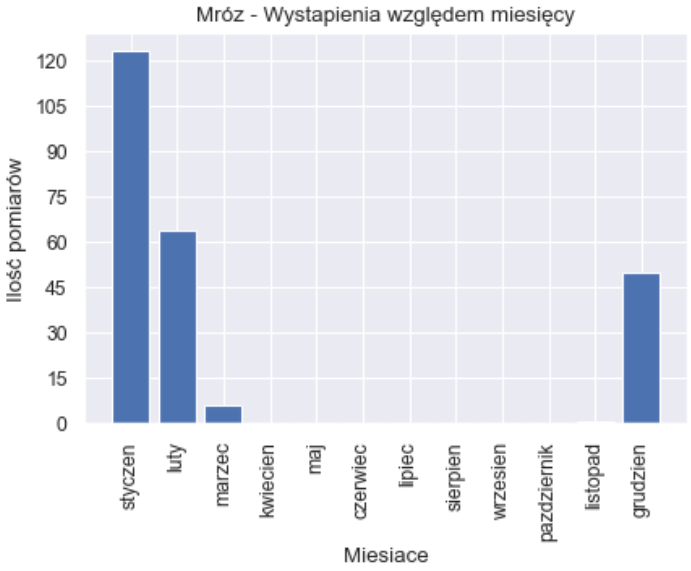
Rysunek 15 Wystąpienia burzy w zależności od średniej temperatury i wilgotności

Ekstremalne temperatury - w przypadku upałów rozpatrywane są dni w których: a) temperatura max przekroczyła 30°C, b) temperatura min przekroczyła 15°C, natomiast mróz gdy: a) temperatura min przyjęła wartość poniżej -15°C, b) temperatura max nie przekroczyła -5°C. Badane są zależności pomiędzy:

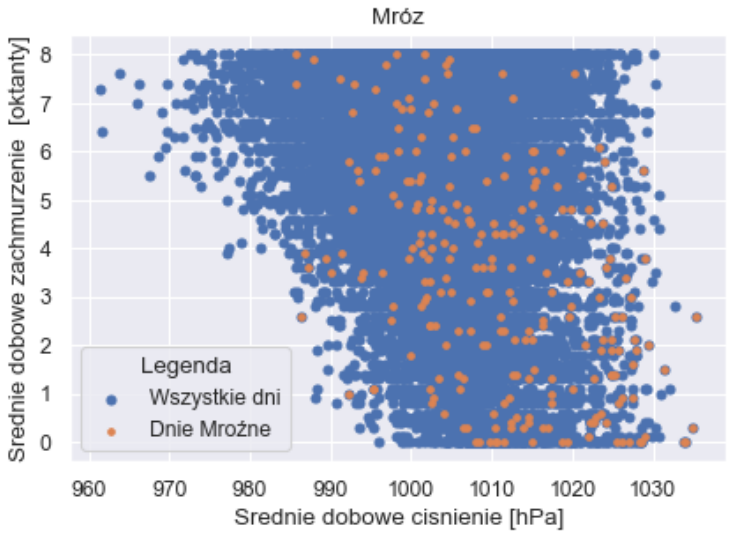
1. Rozkładem ilości wystąpień zjawisk w zależności od roku oraz od miesiąca,
2. Średnim zachmurzeniem a średnim ciśnieniem atmosferycznym.



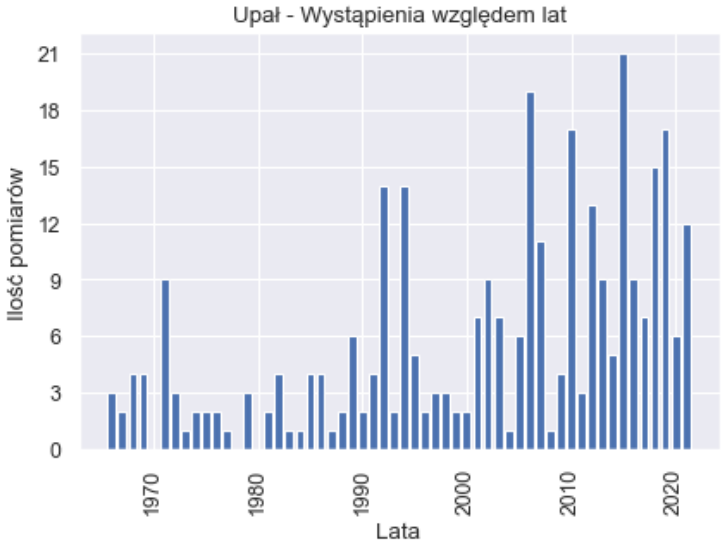
Rysunek 16 Liczba wystąpień mrozów na przestrzeni lat



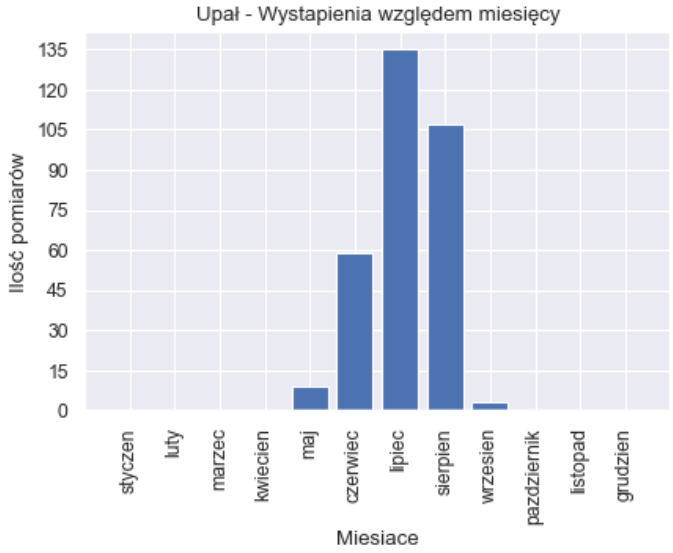
Rysunek 17 Liczba wystąpień mrozów według miesięcy



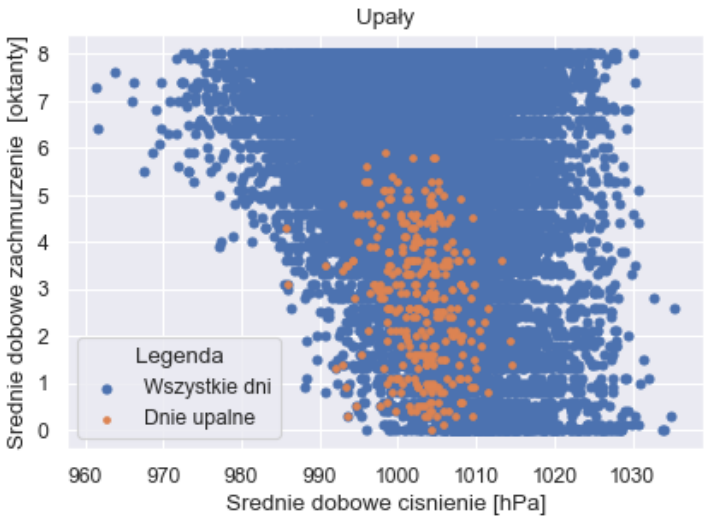
Rysunek 18 Wystąpienia mrozu w zależności od ciśnienia i zachmurzenia



Rysunek 19 Liczba wystąpień upałów na przestrzeni lat



Rysunek 20 Liczba wystąpień upałów według miesięcy



Rysunek 21 Wystąpienia upałów w zależności od ciśnienia i zachmurzenia

Patrząc na powyższe wykresy przedstawiające liczbę wystąpień zjawiska w zadanym okresie czasu można zauważyć, że zgodnie z przypuszczeniami, dane ekstremalne zjawisko, występowało w charakterystycznej dla siebie porze roku(upały i burze dominowały w okresie letnim, ulewy w trakcie lata oraz wczesnej jesieni, śnieżyce i mrozy w okresie zimowym, natomiast wichury nie posiadają charakterystycznej dla siebie pory roku). Natomiast jeżeli chodzi o zmiany na przestrzeni lat, to w przypadku ulew nie nastąpiła wyraźna zmiana, śnieżyc było więcej przed rokiem 1990 niż po (około 2x więcej), w przypadku wichur odnotowany został gwałtowny spadek wystąpień po roku 2010 (możliwa przyczyna to rozwój urbanistyki miasta), burze charakteryzują się stosunkowo stabilną liczbą wystąpień, trudno określić tendencję zmian dla mrozu, natomiast w przypadku upałów widać wyraźny wzrost przypadków po roku 2000.

Patrząc na pozostałe wykresy dla poszczególnych ekstremalnych zjawisk pogodowych można dojść do następujących wniosków:

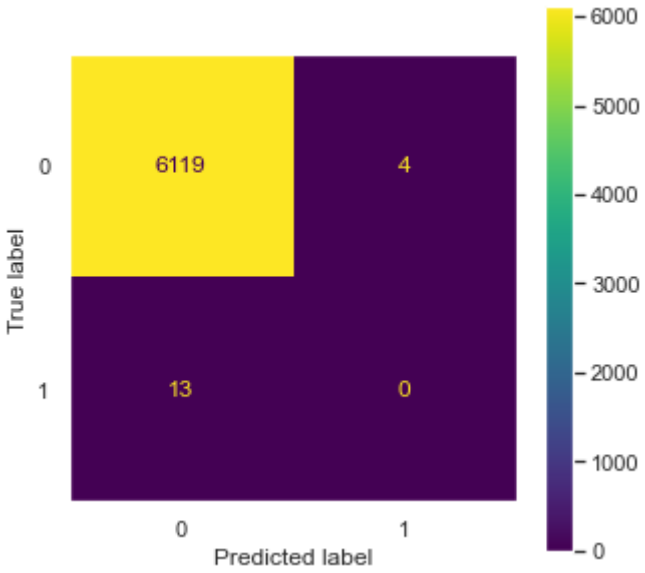
- Ulewa – zjawisko najczęściej występuje w przedziale ciśnień 990-1005hPa, podobnie jak największe deszcze nie klasyfikowane jako ulewa. Większość odnotowanych przypadków wystąpiła dla wilgotności >75% i średniego zachmurzenia >5 oktantów
- Wichura – zjawisko uzależnione jest od występowania lokalnych różnic ciśnienia, o których brak kompletnej informacji w analizowanej bazie danych. W zastępstwie wykorzystana została różnica średnich ciśnień z dwóch kolejnych dni z nadzieją na uzyskanie przydatnych informacji. Zaobserwować można tendencję, że wichury występują gdy pomiędzy kolejnymi dniami odnotowany zostanie spadek ciśnienia, a także przy temperaturze dodatniej.
- Śnieżyca – odnotowywana była najczęściej od -12°C do 2°C. Ze względu na to, że zjawisko śnieżycy zależy od wystąpienia silnego wiatru (podobnie jak wichura) tutaj też pojawił się kłopot z doborem parametru mocno wpływającego na wystąpienie zjawiska.
- Burza – najczęściej obserwowana dla dni o temperaturze średniej przekraczającej 10°C oraz wilgotności >60%, ponadto, wbrew intuicji, burza w okresie zimowym jest zjawiskiem rzadkim, lecz nie nadzwyczajnym.
- Mróz – występuje zazwyczaj przy ciśnieniu wyższym niż 1000hPa, ponadto zachmurzenie nie jest czynnikiem mającym duży wpływ na wystąpienie zjawiska. Ze względu na krótsze dni oraz niski kąt padania promieni słonecznych w zimę, mróz może wystąpić nawet w bezchmurny dzień
- Upał – w odróżnieniu od mrozu, widać wyraźny wpływ zachmurzenia na występowanie zjawiska, przy średnim zachmurzeniu > 6 oktantów nie został odnotowany żaden przypadek przez ostatnie 50 lat. Widoczna jest także zależność od ciśnienia – najczęściej upał występuje pomiędzy 995 – 1010 hPa.

Klasyfikacja z zastosowaniem szeregu czasowego

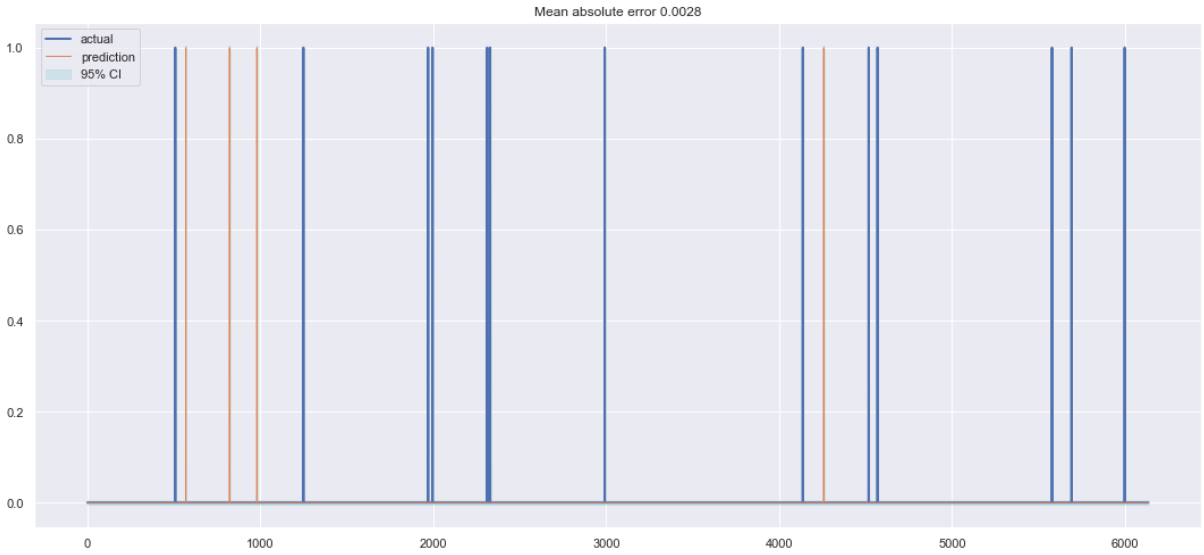
Jest to model predykcji który jako zmienne wykorzystuje zmienne z poprzednich dni. Model ma pozwolić odpowiedzieć na pytanie czy w dniu x wystąpi ekstremalne zjawisko pogodowe, na podstawie danych z dni wcześniejszych (x-1, x-2,…,x-n). Sprawdzane rodzaje klasyfikatorów: Adaboost, RandomForest, ExtraTrees, Bagging. Dla każdego klasyfikatora w danym zjawisku meteorologicznym utworzona została macierz pomyłek w celu oceny pracy modelu. Wielkość stosowanego okna została dobrana w zależności od rozpatrywanego zjawiska meteorologicznego.

Przykładowe wykresy predykcji i tablice pomyłek:

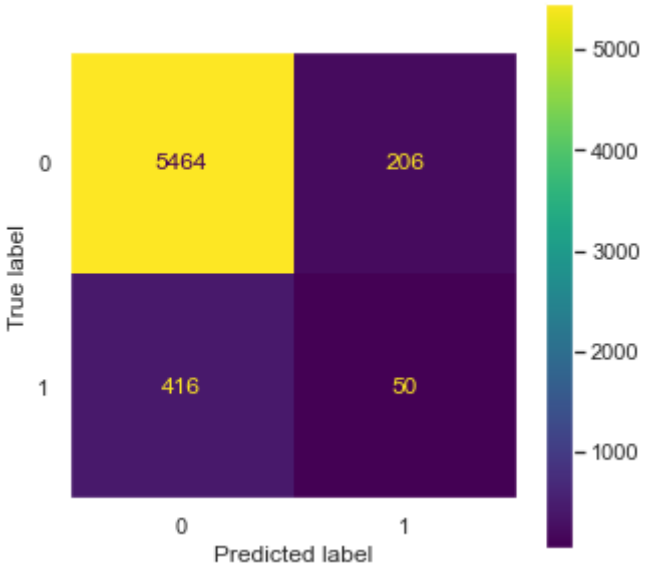
- Ulewa (wielkość okna: 2, rozpatrywane parametry: średnia wilgotność względna, średnie dobowe zachmurzenie)
- Śnieżyca (wielkość okna: 3, rozpatrywane parametry: wysokość pokrywy śnieżnej, średnia dobowa temperatura)
- Wichura (wielkość okna: 2, rozpatrywane parametry: średnie dobowe ciśnienie, średnia dobowa temperatura)
- Burza (wielkość okna: 2, rozpatrywane parametry: średnia wilgotność względna, średnia dobowa temperatura)
- Upał (wielkość okna: 4, rozpatrywane parametry: średnie dobowe zachmurzenie, maksymalna temperatura dobowa, średnia temperatura dobowa)
- Mróz (wielkość okna: 7, rozpatrywane parametry: minimalna dobowa temperatura, maksymalna temperatura dobowa, średnia temperatura dobowa)



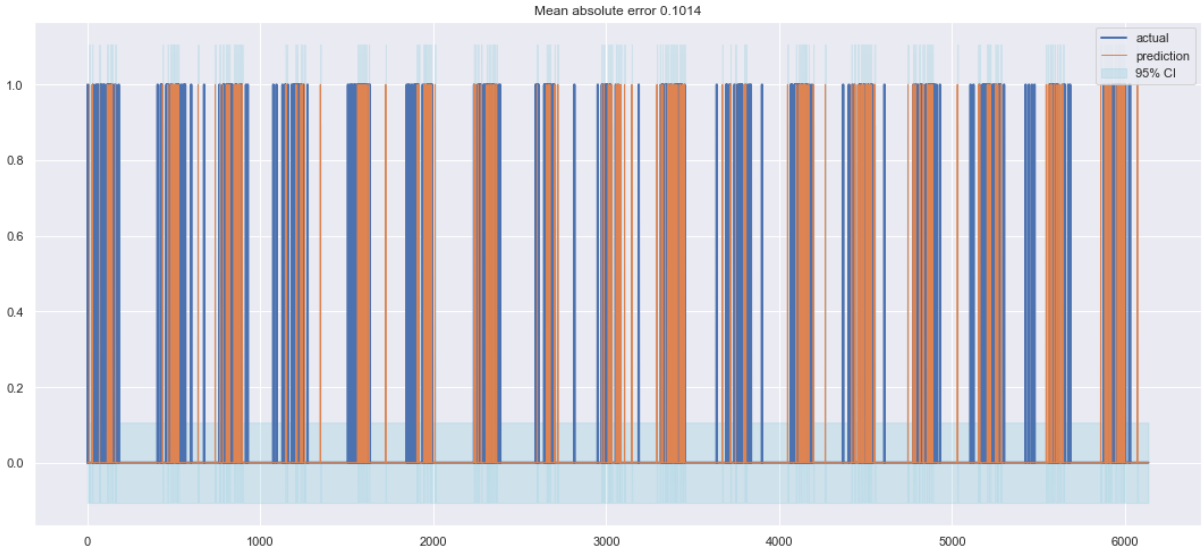
Rysunek 22 Ulewa – BaggingClassificator



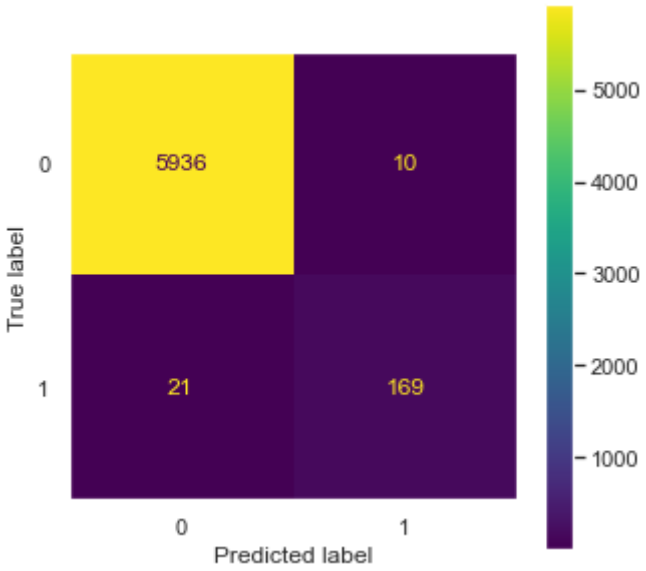
Rysunek 23 Predicted vs Actual (Ulewa - BaggingClassificator)



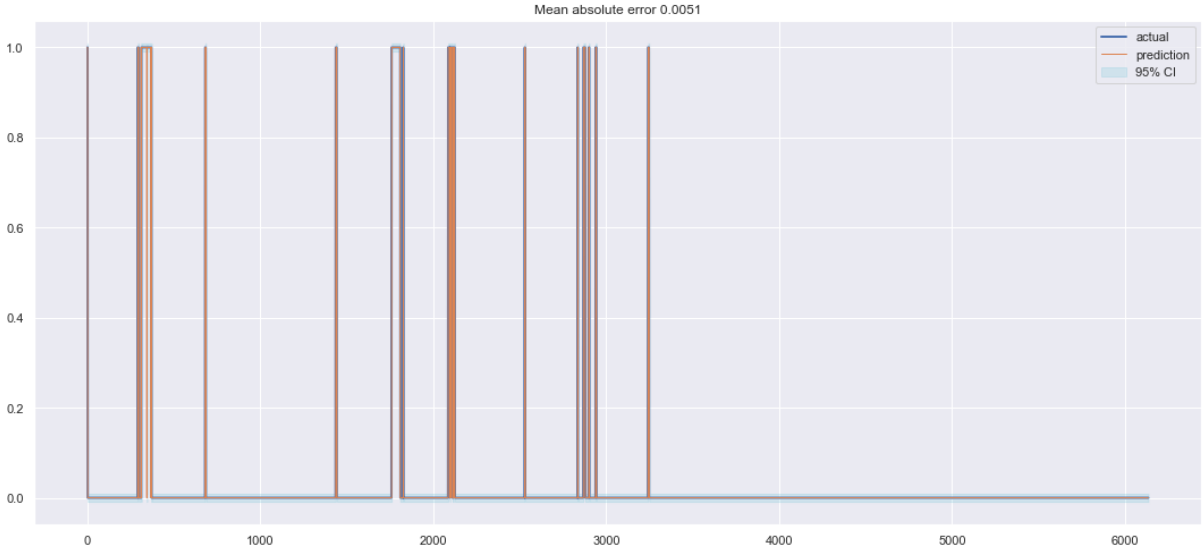
Rysunek 24 Burza - BaggingClassificator



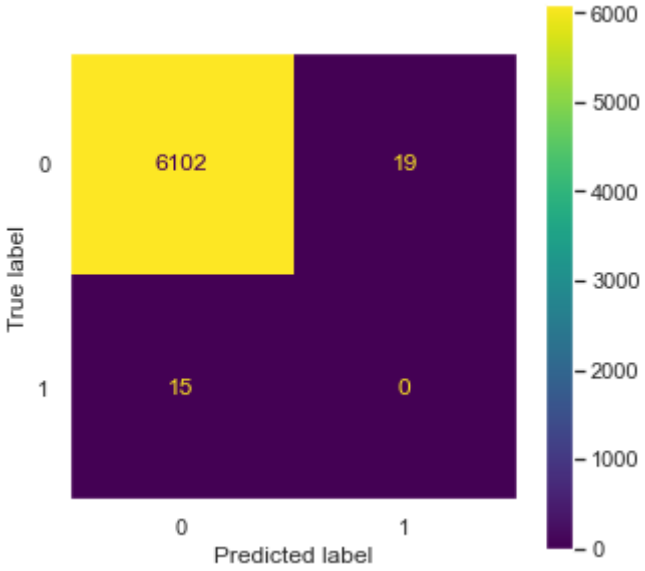
Rysunek 25 Predicted vs Actual (Burza - BaggingClassificator)



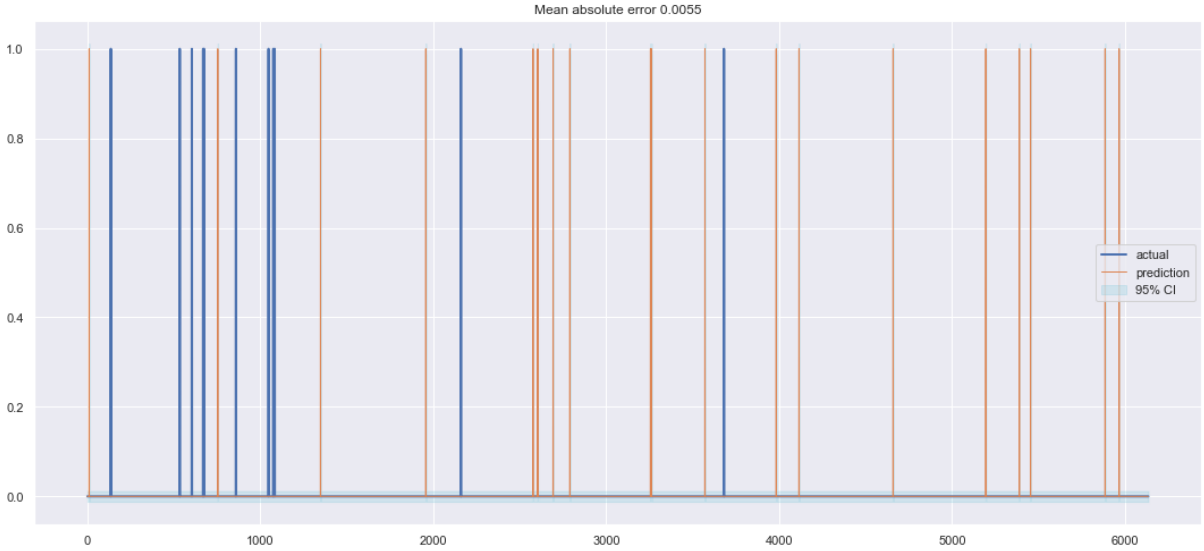
Rysunek 26 Śnieżyca - AdaBoostClassifier



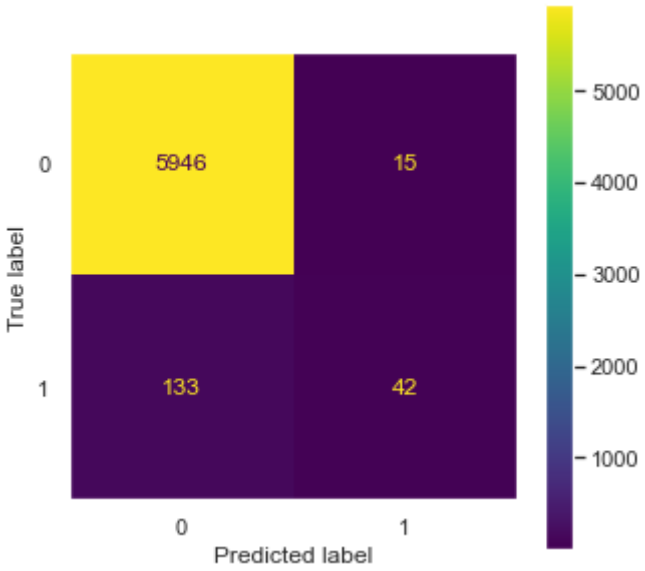
Rysunek 27 Predicted vs Actual (Śnieżyca - AdaBoostClassifier)



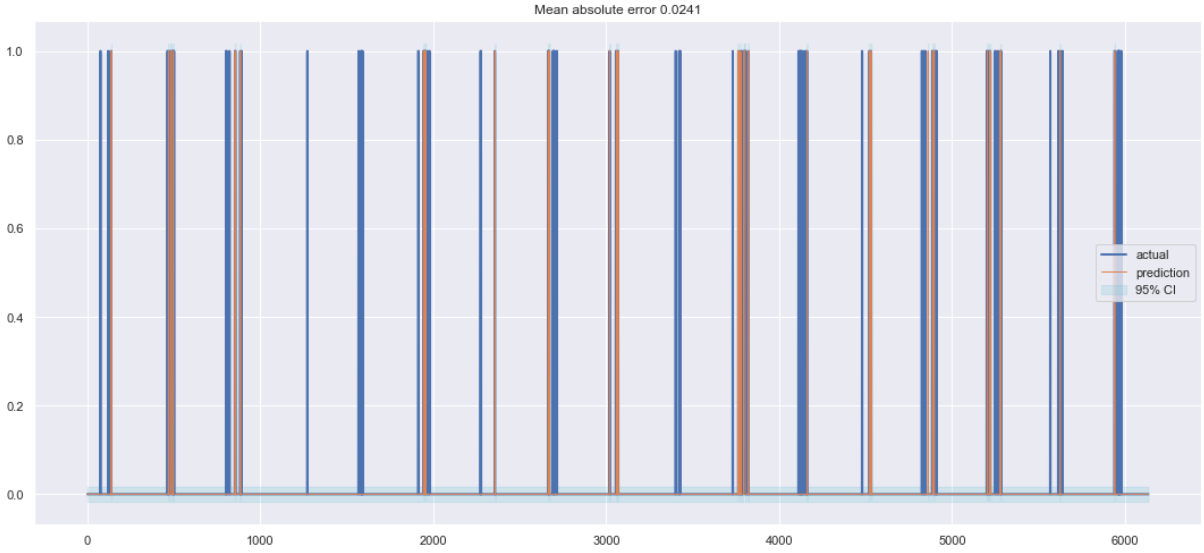
Rysunek 28 Wichura – ExtraTreesClassifier



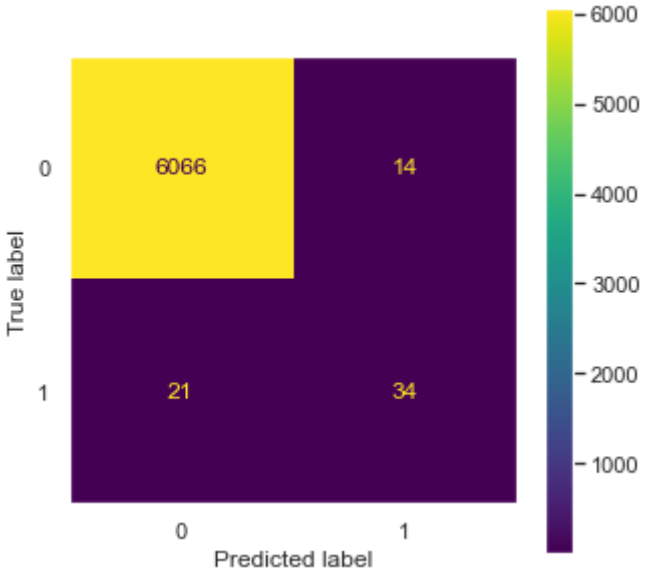
Rysunek 29 Predicted vs Actual (Wichura – ExtraTreesClassifier)



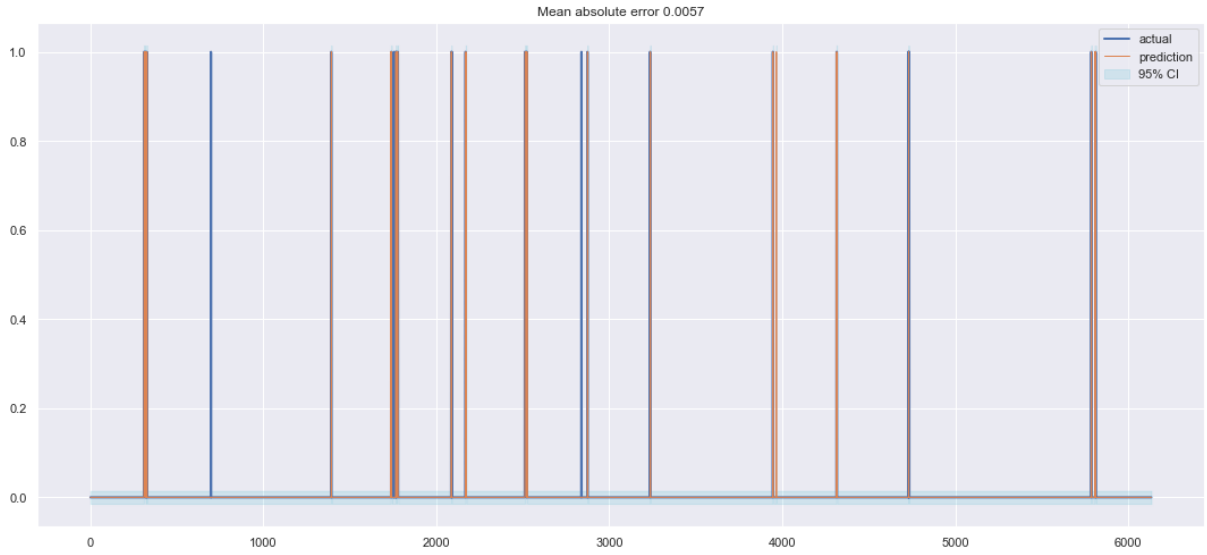
Rysunek 30 Upał - AdaBoostClassifier



Rysunek 31 Predicted vs Actual (Upał - AdaBoostClassifier)



Rysunek 32 Mróz - AdaBoostClassifier



Rysunek 33 Predicted vs Actual (Mróz - AdaBoostClassifier)

W przypadku każdego z zjawisk przetestowane zostały różne klasyfikatory (Adaboost, Randomtrees, Extratrees, Bagging), z zastosowaniem różnego przesunięcia okna serii danych w czasie (mające na celu umożliwienie wykorzystania danych występujących w dniach poprzednich do analizy), dla różnych zestawów danych, wstępnie określonych jako mogące mieć wpływ na zjawisko. Spośród wszystkich przypadków, wyłącznie dla śnieżycy odnotowano zadawalającą efektywność. Jak pokazuje przykład z burzą, pomimo zdecydowanie większej liczby wystąpień, na podstawie dostępnych danych, dalej bardzo trudne (lub wręcz niemożliwe) jest stworzenie klasyfikatora pozwalającego na przewidzenie zjawiska. Można podejrzewać, że sukces w przypadku śnieżycy wynika z tego, że do odnotowania wystąpienia potrzebna jest stosunkowo duża wartość wysokości pokrywy śnieżnej. Ze względu na to, że z jednej strony parametr ten jest mocno uzależniony od temperatury panującej w ciągu całego dnia, i na to że, charakteryzuje się on zdecydowanie większą bezwładnością niż np. ciśnienie, przewidzenie wystąpienia jest dużo łatwiejsze.

Dla wichury i ulewy nie udało się w ogóle zaklasyfikować poprawnie żadnego przypadku dla żadnej sprawdzonej kombinacji parametrów. Wynikać to może, z bardzo dużej dysproporcji między klasami w badanym zbiorze oraz z tego, że pośród używanych danych brakuje takich które odwzorowywałyby

ważne, w ich przypadkach, zjawiska w lokalnej pogodzie takich jak: lokalna różnica ciśnień, gwałtowny spadek temperatury, zderzenie frontów o różnej wilgotności i temperaturze.

W przypadku skrajnych temperatur, lepiej udało się odwzorować mróz ze względu, między innymi na to, że pojawienie się warstwy zalegającego śniegu powoduje wzrost bezwładności temperatury. Ponadto w okresie zimowym ze względu na mały kąt padania światła słonecznego oraz małą długość doby, wzrost temperatury za dnia jest odczuwalnie mniejszy niż w przypadku lata. W przypadku wysokich temperatur w okresie letnim nie pojawia się dodatkowe zjawisko zwiększające bezwładność temperatur.

W przypadku zastosowania za dużego okna, lub ilości danych, został zaobserwowany spadek predykcji, które polegało na tym, że dostarczenia dodatkowych danych do klasyfikatora wpływało w niektórych przypadkach negatywnie na otrzymane ostatecznie wyniki, przede wszystkim prowadziło to do zmniejszenia ilości TP, w trakcie testów regresora.

Uczenie asocjacji

Do badania zjawisk pogodowych została wybrana technika uczenia asocjacji. Uczenie asocjacji zostało wykorzystane w celu odkrycia grup parametrów meteorologicznych występujących najczęściej w momencie obserwacji konkretnego zjawiska pogodowego. Asocjację w rozważanym przypadku można przedstawić przy pomocy pytania: „Jakie wartości parametrów meteorologicznych i w jakiej konfiguracji występują często podczas danego zjawiska pogodowego?”. Podejście to zostało wybrane ze względu na wyznaczoną w ramach poprzedniego etapu, niską korelację pomiędzy czynnikami meteorologicznymi, a wystąpieniem zjawiska ekstremalnego. Do uczenia asocjacji wykorzystano jeden z najpowszechniejszych algorytmów – algorytm Apriori. W tym celu należało specjalnie przygotować do tego bazę danych, tak aby analizie poddany został tylko jej fragment, zawierający interesujące elementy. Dla każdego z badanych czynników stworzone zostały kolumny reprezentujące pewne przedziały wartości. W zależności od wartości parametru odpowiednie pola w stworzonych kolumnach ustawiane były na True. Wstępnie dane zostały podzielone na przedziały o równych szerokościach (np. dla średniego dobowego zachmurzenia ogólnego, przyjmuje się wartości rzeczywiste od 0 do 8, więc utworzone zostało 8 kolumn reprezentujących różne przedziały z których w konkretnym dniu tylko w jednej występuje wartość True). Następnie dla badanych zjawisk meteorologicznych (ulewa, śnieżyca, wichura, burza, upał oraz mróz) zostały utworzone zbiory zawierające dane wybrane w poprzednim etapie, dla których wykonane zostało poszukiwanie reguł asocjacji z wykorzystaniem algorytmu apriori:

- Antecedents - pierwsze zjawisko/zjawiska
- Consequents - następne zjawisko/zjawiska
- Antecedent support- prawdopodobieństwo wystąpienia antecedents
- Consequents support- prawdopodobieństwo wystąpienia consequents
- Support- prawdopodobieństwo wystąpienia antecedents i consequents jednocześnie
- Confidence- prawdopodobieństwo wystąpienia consequents jeśli wystąpiło antecedents
- Lift- kiedy wystąpił antecedents, prawdopodobieństwo wystąpienia consequents wzrasta o wartość lift
- Leverage- podobny do lift ale w naszym problemie będziemy go pomijać
- Conviction- szans na wystąpienie antecedents bez consequents
- Length- ilość elementów antecedents

Do uczenia asocjacji dla każdego zjawiska pogodowego wykorzystane zostały takie same parametry jak przy eksploracyjnej analizie danych oraz dodana została nowa kolumna zawierająca informacje o porze roku (przyjęto wiosnę od 1 marca do 31 maja, lato od 1 czerwca do 31 sierpnia, jesień od 1 września do 30 listopada, zima od 1 grudnia do 28/29 lutego). Oryginalny zbiór wszystkich

dni został pomniejszony tylko do dni, w których można było zaobserwować analizowane zjawisko meteorologiczne. Następnie zmieniono sposób podziału przedziałów z równej wartości na taką samą liczbę próbek w celu weryfikacji, która z metod jest lepsza dla naszego przypadku, a która powoduje za dużą stratę informacji w wyniku postawionych granic przedziałów.

Poniżej przedstawiono przykładowe wyniki działania algorytmu apriori przy różnych parametrach min_support, dla przedziałów o równej szerokości (same_width) lub równej liczbie próbek (same_amount). Wybrano te powiązania które w kontekście rozwiązywanego problemu wydają się interesujące i mogą pomóc w predykcji ekstremalnych zjawisk pogodowych. W przypadku same_width przedziały nazywane są od nazwy parametru i wartości przez nie przyjmowane, tak więc gdy wilgotność nazywa się Wilgotność(80,90) chodzi o zakres wilgotności do 80% do 90%.

Ulewa:

Tabela 3 Reguły asocjacji dla ulewy przy parametrach: Min_supp = 50%, same_width (wszystkie wyniki)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	length
'Lato'	'Ulewa'	0.761904762	1	0.761905	1	1	0	inf	1
'Temp(10,20)'	'Ulewa'	0.619047619	1	0.619048	1	1	0	inf	1
'Wilgotnosc(90,100)'	'Ulewa'	0.5	1	0.5	1	1	0	inf	1
'Cisnienie(990,1000)'	'Ulewa'	0.595238095	1	0.595238	1	1	0	inf	1

Tabela 4 Reguły asocjacji dla ulewy przy parametrach: Min_supp = 50%, same_amount (wszystkie wyniki)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	length
'Lato'	'Ulewa'	0.761904762	1	0.761905	1	1	0	inf	1

Tabela 5 Reguły asocjacji dla ulewy przy parametrach: Min_supp = 25%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	consequent supp	support	confidence	lift	leverage	conviction	leng
'Temp(10,20)', 'Wilgotnosc(90,100)'	'Ulewa'	0.404761905	1	0.4048	1	1	0	inf	2
'Temp(10,20)', 'Cisnienie(990,1000)'	'Ulewa'	0.404761905	1	0.4048	1	1	0	inf	2
'Temp(10,20)', 'Wilgotnosc(90,100)'	'Ulewa', 'Lato'	0.404761905	0.761904762	0.3333	0.823529	1.1	0.0249	1.349206	2
'Temp(10,20)', 'Cisnienie(990,1000)'	'Ulewa', 'Lato'	0.404761905	0.761904762	0.3333	0.823529	1.1	0.0249	1.349206	2
Temp(10,20)', 'Cis(990,1000)', 'Wilg.(90,100)'	'Ulewa'	0.261904762	1	0.2619	1	1	0	inf	3

Tabela 6 Reguły asocjacji dla ulewy przy parametrach: Min_supp = 25%, same_amount (wybrane wyniki)

antecedents	consequents	antecedent suppo	consequent suppo	suppo	confider	lift	levera	convicti	leng
'Temp<12.825'	'Ulewa'	0.261904762	1	0.2619	1	1	0	inf	1
'Temp>21.15'	'Ulewa'	0.261904762	1	0.2619	1	1	0	inf	1
'Temp>21.15'	'Lato'	0.261904762	0.761904762	0.2619	1	1.3125	0.06236	inf	1
'Ulewa', 'Lato'	'Temp>21.15'	0.761904762	0.261904762	0.2619	0.34375	1.3125	0.06236	1.1247166	2
'Lato'	'Ulewa', 'Temp>21.15'	0.761904762	0.261904762	0.2619	0.34375	1.3125	0.06236	1.1247166	1

Wykorzystując metodę przedziałów same_width uzyskano więcej wyników o większym supporcie niż przy metodzie same_amount. Wynika to ze specyfiki podziału na równoliczne podzbiory. Podejście to sprawia, że dla obszarów o dużym zagęszczeniu próbek wyznaczone zostają bardzo wąskie przedziały. Ze względu na sposób obliczania supportu na tak stworzonym zbiorze algorytm apriori zadziała niezgodnie z oczekiwaniami. Uzyskanie wysokiego supportu będzie niemożliwe. Dla przykładu dla Ulewa przedziały wilgotności dla same_width to: 0-10%, 11-20%,…,91-100%, natomiast dla same_amount przedziały wilgotności przyjmują następujące wartości: 0-74.47%, 74.47-79.46%, 79.46-83.07%, 83.07-86.18%, 86.18-90.75%, 90.75-93.06%, 93.06-94.68%, 94.68-95.42%, 95.42-96.82%, 96.82-100%. Dla wilgotności dominują wartości z przedziału 75-100%, natomiast przy zastosowaniu podejścia same_amount jest to reguła niemożliwa do zaobserwowania przez algorytm apriori. Wniosek ten skłonił nas do uznania metody same_amount niewłaściwej dla rozpatrywanego problemu.

Śnieżyca:

Tabela 7 Reguły asocjacji dla śnieżycy przy parametrach: Min_supp = 50%, same_width (wybrane wyniki)

antecedents	consequents	antecedent support	cons. support	support	confidence	lift	convic	len
'Zima'	'Sniezyca'	0.948717949	1	0.948718	1	1	inf	1
'Wiatr>10', 'Opad(0,10)', 'Temp(-10,0)'	'Zima'	0.717948718	0.948717949	0.692308	0.9642857	1.016	1.436	3
'Wiatr>10', 'Opad(0,10)', 'Temp(-10,0)'	'Sniezyca'	0.717948718	1	0.717949	1	1	inf	3
'Zima', 'Wiatr>10', 'Opad(0,10)', 'Temp(-10,0)'	'Sniezyca'	0.692307692	1	0.692308	1	1	inf	4

Tabela 8 Reguły asocjacji dla śnieżycy przy parametrach: Min_supp = 25%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	cons. supp	support	confiden	lift	convic	len
'Wiatr>10', 'Opad(0,10)', 'Zima', 'Wilgotnosc(80,90)', 'Temp(-10,0)'	'Sniezyca'	0.358974359	1	0.35897	1	1	inf	5
'Wiatr>10', 'Cisnienie(990,1000)', 'Opad(0,10)', 'Zima', 'Temp(-10,0)'	'Sniezyca'	0.282051282	1	0.28205	1	1	inf	5
'Wiatr>10', 'Opad(0,10)', 'Zima', 'Pokrywa(20,30)', 'Temp(-10,0)'	'Sniezyca'	0.307692308	1	0.30769	1	1	inf	5
'Wilgotnosc(90,100)', 'Wiatr>10', 'Opad(0,10)'	'Zima', 'Sniezyca'	0.333333333	0.9487179	0.30769	0.92308	1	0.667	3

Wichura:

Tabela 9 Reguły asocjacji dla wichury przy parametrach: Min_supp = 50%, same_width (wszystkie wyniki)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	conviction	lenght
'Temp(0,10)'	'Wichura'	0.700854701	1	0.7008547	1	1	inf	1
'dP(-10,0)'	'Wichura'	0.555555556	1	0.5555556	1	1	inf	1

Tabela 10 Reguły asocjacji dla wichury przy parametrach: Min_supp = 25%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	consequent supp	support	confidence	lift	conviction	lenght
'dP(-10,0)', 'Temp(0,10)'	'Wichura'	0.41025641	1	0.4102564	1	1	inf	2
'Wilgotnosc(80,90)'	'Wichura'	0.487179487	1	0.4871795	1	1	inf	1
'Temp(0,10)', 'Zima'	'Wichura'	0.324786325	1	0.3247863	1	1	inf	2
'Temp(0,10)', 'Zachmurzenie(6,7)'	'Wichura'	0.282051282	1	0.2820513	1	1	inf	2
'Temp(0,10)', 'Wilgotnosc(80,90)'	'Wichura'	0.376068376	1	0.3760684	1	1	inf	2
'Temp(0,10)', 'Wilgotnosc(70,80)'	'Wichura'	0.256410256	1	0.2564103	1	1	inf	2

Burza:

Tabela 11 Reguły asocjacji dla burzy przy parametrach: Min_support = 50%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	consequent supp	support	confiden	lift	conviction	length
'Lato'	'Burza'	0.648411089	1	0.648411	1	1	inf	1
'Opad(0,10)'	'Burza'	0.797836376	1	0.797836	1	1	inf	1
'Temp(10,20)'	'Burza'	0.624070318	1	0.62407	1	1	inf	1
'Srwiatr(2,4)'	'Burza'	0.607843137	1	0.607843	1	1	inf	1

Tabela 12 Reguły asocjacji dla burzy przy parametrach: Min_support = 25%, same_width (wybrane wyniki)

antecedents	consequ	antecedent supp	consequent supp	support	conf	lift	convict	len
'Lato', 'Opad(0,10)'	'Burza'	0.492900609	1	0.492901	1	1	inf	2
'Temp(10,20)', 'Cisnienie(990,1000)'	'Burza'	0.329952671	1	0.329953	1	1	inf	2
'Temp(10,20)', 'Cisnienie(1000,1010)'	'Burza'	0.258958756	1	0.258959	1	1	inf	2
'Lato', 'Srwiatr(2,4)', 'Opad(0,10)'	'Burza'	0.320486815	1	0.320487	1	1	inf	3
'Temp(10,20)', 'Srwiatr(2,4)', 'Opad(0,10)'	'Burza'	0.288708587	1	0.288709	1	1	inf	3

Upał:

Tabela 13 Reguły asocjacji dla upału przy parametrach: Min_support = 50%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	consequent supp	support	confidence	lift	conviction	len
'Cisnienie(1000,1010)', 'Tempmax(30,40)'	'Lato'	0.738019169	0.961661342	0.722045	0.978355	1.0174	1.771246	2
'Tempmax(30,40)', 'Srwiatr(2,4)', 'Lato'	'Upal'	0.632587859	1	0.632588	1	1	inf	3
'Cisnienie(1000,1010)', 'Tempmax(30,40)', 'Lato'	'Upal'	0.722044728	1	0.722045	1	1	inf	3

Tabela 14 Reguły asocjacji dla upału przy parametrach: Min_support = 25%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	consequent supp	support	conf	lift	conv	length
'Cisnienie(1000,1010)', 'Tempmax(30,40)', 'Srwiatr(2,4)', 'Lato'	'Upal'	0.485623003	1	0.485623	1	1	inf	4
'Cisnienie(1000,1010)', 'Tempmax(30,40)', 'Srwiatr(2,4)'	'Upal'	0.498402556	1	0.498403	1	1	inf	3

Mróz:

Tabela 15 Reguły asocjacji dla mrozu przy parametrach: Min_support = 50%, same_width (wybrane wyniki)

antecedents	consequents	antecedent supp	consequent supp	support	confiden	lift	convict	len
'Mroz'	'Zima'	1	0.971311475	0.971311	0.9713	1	1	1
'Zima'	'Mroz'	0.971311475	1	0.971311	1	1	inf	1
'Mroz'	'Tempmin(-20,-10)'	1	0.692622951	0.692623	0.6926	1	1	1
'Tempmin(-20,-10)'	'Zima'	0.692622951	0.971311475	0.668033	0.9645	0.99	0.808	1
'Tempmin(-20,-10)', 'Zima'	'Mroz'	0.668032787	1	0.668033	1	1	inf	2

Tabela 16 Reguły asocjacji dla mrozu przy parametrach: Min_support = 25%, same_width (wybrane wyniki)

antecedents	consequents	ant. supp.	cons. supp.	support	conf.	lift	convict	len
'Tempmin(-20,-10)', 'Srwiatr(2,4)'	'Mroz'	0.348361	1	0.348	1	1	inf	2
'Tempmin(-30,-20)'	'Mroz', 'Zima'	0.303279	0.9713115	0.299	0.99	1.02	2.123	1
'Tempmin(-20,-10)', 'Srwiatr(2,4)'	'Mroz', 'Zima'	0.348361	0.9713115	0.332	0.95	0.98	0.61	2
'Tempmin(-20,-10)'	'Srwiatr(2,4)'	0.692623	0.4508197	0.348	0.5	1.12	1.105	1
'Zima'	'Srwiatr(2,4)'	0.971311	0.4508197	0.43	0.44	0.98	0.986	1
'Zima'	'Tempmin(-20,-10)', 'Mroz', 'Srwiatr(2,4)'	0.971311	0.3483607	0.332	0.34	0.98	0.99	1

Dla powyższych zjawisk zaobserwowano kilka interesujących, pod kątem predykcji ekstremalnych zjawisk pogodowych, zestawów parametrów. Warto zaznaczyć na samym wstępie, że w momencie gdy w kolumnie consequents występuje tylko ekstremalne zjawisko to confidence zawsze będzie równy 100%. Wynika to z faktu, że przyjęty zbiór danych zawiera tylko dni gdy owe zjawisko wystąpiło, dlatego w tym przypadku ważną metryką dla całego modelu będzie support. Natomiast gdy rozpatrywane będą wszystkie dni ze zbioru, to confidence będzie bardzo ważnym parametrem określającym dobroć modelu, ponieważ będzie określać prawdopodobieństwo wystąpienia zjawiska meteorologicznego jeśli wystąpiły dane parametry. Podobną sytuację można zaobserwować w przypadku conviction, gdzie widnieje wartość inf (infinit). Spowodowane jest to tym że dla aktualnego zbioru nigdy nie wystąpi zależność podana w antecedents bez wystąpienie consequents.

W przypadku ulewy najciekawsza zależność to [Wilgotność(90 - 100%), Ciśnienie(990 - 1000hPa), Temperatura(10 - 20C)] - > Ulewa. Pokazuje to, że ponad 26% ulew w ostatnich 60 latach miało wilgotność w przedziale 80-90%, ciśnienie 990-1000 hPa i temperaturę 10-20 ^{o}C (co potwierdzają wykresy: rysunek 3 i 4). Również warto zauważyć, że [Lato] - > Ulewa pokazuje, że około 76% wszystkich ulew przypadło w okresie czerwiec-sierpień. Pozostałe zależności o wysokim supporcie to różne kombinacje wcześniej wymienionych parametrów. Podobną zależność będzie można zaobserwować w pozostałych zjawiskach meteorologicznych.

Dla śnieżyc interesującą zależnością jest [Zima, Wiatr>10, Opad(0,10), Temp(-10,0)] -> Śnieżyca, która wystąpiła w około 69% przypadków. Również zależność [Wiatr>10, Opad(0,10), Temp(-10,0)] -> Śnieżyca z suportem na poziomie 72% pokazuje, że dane warunki wraz z zjawiskiem meteorologicznym występowały w innych porach roku niż zima. Warto zauważyć 3 przypadki szeregu pięciu warunków atmosferycznych, w których wystąpiła śnieżyca. Zależności [Wiatr>10, Opad(0,10), Zima, Temp(-10,0), Wilgotność(80,90)/Ciśnienie(990,1000)/Pokrywa(20,30)] -> Śnieżyca osiągają kolejno 36%, 28% oraz 31%. Pomimo dzielenia ze sobą 4 identycznych parametrów nie wykryto asocjacji zawierającej wszystkie te parametry (szereg 7 warunków ma niskie prawdopodobieństwo spełnienia się).

Wichura posiada jedynie 2 zależności przekraczające 50%: [Temp(0,10)] -> Wichura oraz [dP(-10,0)] -> Wichura, symbolem d oznaczono różnicę parametru (w tym przypadku ciśnienia) z sąsiadujących dni. Więcej informacji można wywnioskować z drugiej tablicy gdzie min_supp = 25%. Od razu rzuca się w oczy fakt, że prawie połowa wichur występowała dla wilgotności z przedziału 80-90%. Zgodnie z intuicją, wichurze najczęściej towarzyszy duże zachmurzenie, [Temp(0,10), Zachmurzenie(6,7)] -> Wichura o wartości 28%. Pozostałe przedziały zachmurzeń nie przekroczyły 25% supportu.

W przypadku burz, podobnie jak ulewy, największa liczba zjawisk miała miejsce latem (ok 65% wszystkich przypadków). Interesującą zależnością jest [Srwiatr(2,4), Temp(10,20), Opad(0,10)] -> Burza, która wystąpiła w około 29% burz. Kolorem zaznaczono zależności wyznaczone przez algorytm apriori w przypadku których, pomimo stosunkowo wysokiego supportu, trudno mówić o wynikaniu np. [Opad(0,10), Lato] -> Burza ponieważ takie warunki mogą występować dużo częściej bez konieczności wystąpienia burzy i w momencie gdy przełączymy zbiór z tylko dni burzowych na wszystkie dni support i confidence zmaleją.

Przedostatnim rozpatrywanym zjawiskiem meteorologicznym jest upał. Zależność [Ciśnienie(1000,1010), Tempmax(30,40), Lato] -> Upał sprawdza się w ponad 72% wszystkich upałów. Zgodnie z przypuszczeniami, niewielka średnia dobowa prędkość wiatru sprzyja występowaniu upałów. Podobnie jak w przypadku śnieżycy tutaj również można zaobserwować negatywny wpływ pory roku na support (tabela 14) informujący nas o tym, że upał wystąpił dla podanych wartości parametrów w innym okresie.

Na koniec przejdźmy do mrozu, który zgodnie z założeniami w większości dni przypada na zimę [Zima] -> Mróz około 97% (należy pamiętać, że przyjęta przez nas zima nie jest zgodna z astronomiczną zimą i nie obejmuje miesiąca marzec). [Zima] -> Srwiatr(2,4) to przykład wątpliwego wynikania, który nie wnosi żadnej informacji dla naszego problemu.

Podsumowanie i wnioski

Na bazie zbioru w którym występują silne dysproporcje między ilością danych a obserwacjami interesujących zdarzeń możliwe jest poszukiwanie i wyznaczanie pewnych zależności, czy poszukiwanie przyczyn i warunków prowadzących do zaistnienia badanego zjawiska. Nie ma gwarancji, że uda się zawsze coś takiego wyznaczyć lub znaleźć. Tak jak w przypadku analizowanej w ramach projektu wichury, wraz ze zdobywaniem wiedzy dziedzinowej można na pewnym etapie dojść do wniosku, że dane dostępne i wykorzystywane nie pozwalają na bezpośrednie zdobycie informacji o wystąpieniu zjawiska lub okoliczności mu sprzyjających. W skrajnym przypadku możliwe jest też, że brak takiej informacji spowoduje reakcję łańcuchową i uniemożliwi opisanie innych zjawisk ze względu na bezpośrednią, od poprzednich, zależność. Niemniej jednak można czasem odkryć pewne zależności które mogą wydawać się nieoczywiste lub wręcz bezpośrednio zaprzeczać intuicji. W przypadku analizowanej bazy danych i rozpatrywanego problemu do takich obserwacji zaliczyć można zależności opisane przy okazji analizy wyników uzyskanych przy pomocy algorytmu apriori.

Pogoda jest mocno uzależniona od chwilowych zmian parametrów takich jak ciśnienie, czy wilgotność powietrza, duży wpływ ma także odpowiedni układ tych parametrów w przestrzeni, np. powstanie wiatru gwarantuje różnica ciśnień między dwoma miejscami, a powstanie burzy zależy od zderzenia się dwóch frontów powietrza o różnych właściwościach. Ze względu na to, że wykorzystana w ramach projektu baza danych zawierała w większości dane uśrednione na przestrzeni całego dnia, zebrane w jednej lokalizacji, odwzorowanie niektórych zjawisk było bardzo trudne, a znalezione

zależności w wielu przypadkach nie były dokładnym odwzorowaniem przyczyn powstania. Prawdopodobnie, gdyby posiadać więcej przykładów wystąpień skrajnych warunków pogodowych oraz bardziej dokładne dane, np. zmiany ciśnienia czy wilgotności z godziny na godzinę, udałoby się znaleźć lepsze zależności, a przynajmniej lepiej uzasadnić ich występowanie (ponieważ część ekstremalnych zjawisk pogodowych jest zdarzeniem chwilowym i gwałtownym, trwającym maksymalnie kilka godzin). Dodatkową poprawę przyniosłoby też zebranie razem informacji z różnych miejsc w przestrzeni, np. w przypadku analizowanej Warszawy, przeprowadzić pomiary w każdej z dzielnic lub też na przedmieściach, tak aby uzyskać pewnego rodzaju mapę parametrów dla bezpośredniego sąsiedztwa. Dodatkowe informacje na temat lokalnych frontów, w odpowiednio większym dystansie, prawdopodobnie też przyczyniłyby się do zwiększenia skuteczności predykcji. Jednakże samo zdobycie większej bazy danych nie spowodowałoby automatycznie poprawy wyników, należałoby by jeszcze wybrać i przeanalizować odpowiednie zależności. Z drugiej strony pojawiłby się dodatkowy problem jakim byłoby uwzględnienie w odpowiedni sposób rozkładu parametrów w przestrzeni. W przypadku przeprowadzonego projektu, ze względu na pochodzenie danych z jednej stacji pomiarowej, taki problem nie zaistniał.

Pomimo tego, że zjawiska pogodowe wynikają z pewnych ogólnych zależności, przedstawiona analiza i modele są mocno związane z lokalizacją z której pochodziły dane. W zależności od ukształtowania terenu, położenia, sąsiedztwa zbiorników wodnych, czy nawet lokalnej zabudowy, może się okazać, że przygotowany model predykcji dla innego miejsca nie będzie w żadnym stopniu adekwatny.