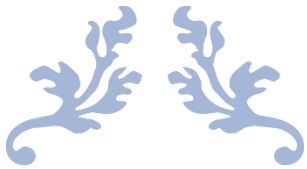


KLASYFIKACJA I PROGNOZA DNI DESZCZOWYCH NA PODSTAWIE CZYNNIKÓW ATMOSFERYCZNYCH Z DNIA POPRZEDNIEGO

Łukasz Wójcik 185173
Michał Senderski 205706
Michał Strus 205939
Paweł Białobrzeski 208509

1. Spis treści

2. Streszczenie

Niniejszy dokument zawiera opis badania pogody na podstawie danych australijskiej agencji meteorologicznej (Bureau of Meteorology) z lat 2008-2017 oraz opracowania algorytmu pozwalającego przewidzieć czy następnego dnia będzie padać przy użyciu różnych algorytmów machine-learningowych. Wśród zastosowanych metod znalazły się lasy losowe, wzmocnienie gradientowe, metoda knn (k najbliższych sąsiadów), regresja logistyczna i SGD oraz metody hybrydowe. Rezultaty uzyskane przez te metody zostały porównane przy pomocy miernika dokładności (accurancy) oraz krzywej ROC. Dodatkowo, aby zminimalizować wpływ losowości podziału zbioru danych na zbiór testowy i uczący zastosowano walidację krzyżową. Ponadto przetestowano działanie algorytmów na sztucznie wprowadzonych obserwacjach.

3. Słowa kluczowe

Klasyfikacja, wzmacnianie gradientowe, lasy losowe, meteorologia, przewidywanie opadów, walidacja krzyżowa, opady deszczu w Australii, przetrenowanie, GridSearch, boxplot, krzywa ROC, miernik AUC, accuracy, macierz konfuzji

4. Wprowadzenie

W ostatnich latach rozwój technologii oraz dostępność rozbudowanych zbiorów danych meteorologicznych stworzyły nowe perspektywy w dziedzinie przewidywania opadów atmosferycznych. Opady deszczu mają istotny wpływ na różnorodne dziedziny, od rolnictwa po zarządzanie zasobami wodnymi, a także stanowią kluczowy element w prognozowaniu katastrof naturalnych. W tym kontekście, badania naukowe skupiające się na wykorzystaniu zaawansowanych algorytmów machine learningowych do klasyfikacji oraz prognozowania opadów następnego dnia stają się coraz bardziej istotne.

Tradycyjne metody prognozowania opadów opierają się na analizie historycznych danych meteorologicznych, jednak w obliczu złożoności i nieliniowości współczesnych zjawisk atmosferycznych, podejście to może okazać się niewystarczające. W odpowiedzi na te wyzwania, algorytmy machine learningowe, takie jak modele klasyfikacyjne zastosowane w naszych badaniach, stanowią obiecujące narzędzie do poprawy precyzji prognoz opadów.

Badania wykorzystujące machine learning w kontekście prognozowania opadów skupiają się na efektywnym wykorzystaniu różnorodnych danych meteorologicznych, takich jak temperatura,

wilgotność, prędkość wiatru czy ciśnienie atmosferyczne. Modele klasyfikacyjne, w tym popularne algorytmy jak Random Forest (lasy losowe) czy Gradient Boosting, są zdolne do analizy tych danych i identyfikacji wzorców, które mogą wskazywać na prawdopodobieństwo wystąpienia opadów w kolejnym dniu, często niestandardowych i niedostrzegalnych przez człowieka.

5. Przedmiot badania

5.1 Cel

Celem niniejszych badań jest zbadanie skuteczności algorytmów machine learningowych w klasyfikacji opadów atmosferycznych na podstawie danych meteorologicznych australijskiego Bureau of Meteorology. Poprzez analizę dużego zbioru danych (dane dzienne z 9 lat), uwzględniającego różnorodne czynniki atmosferyczne, dążymy do opracowania modeli prognozujących, które nie tylko zwiększą precyzję przewidywań, ale także dostarczą bardziej kompleksowego zrozumienia dynamiki opadów. Wyniki tych badań mogą znaleźć zastosowanie w doskonaleniu systemów monitorowania pogody oraz w usprawnianiu działań związanych z zarządzaniem ryzykiem związanym z opadami atmosferycznymi.

5.2 Wstępna analiza danych

a) Przedstawienie dostępnych zmiennych

Na zmienne objaśniające wstępnie wybrano wszystkie zmienne dostępne w wybranych datasecie, z wyjątkiem RainTommorow która jest naszą zmienną objaśnianą (zbiorem etykiet klas). Po analizie do prawdziwego modelu zostały zmienne:

- MinTemp - minimalna zaobserwowana temperatura w przeciągu 24 godzin do 9 rano badanego dnia w stopniach Celsjusza.
- MaxTemp – maxymalna zaobserwowana temperatura w przeciągu 24 godzin do 9 rano badanego dnia w stopniach Celsjusza.
- Rainfall – opady w przeciągu 24 h do 9 rano badanego dnia, mierzone w milimetrach
- Evaporation – zmierzone parowanie wody w standardowym urządzeniu Klasy A, w przeciągu 24 godzin do 9 rano badanego, dnia mierzone w milimetrach.
- Sunshine – czas jasnego słońca w przeciągu 24 do północy, mierzony w godzinach.
- WindGustSpeed - prędkość najszybszego wiatru w przeciągu 24 do północy, mierzony w kilometrach na godzinę.
- WindSpeed9am - uśredniona prędkość z 10 min przed 9 rano badanego dnia, mierzona w kilometrach na godzinę.

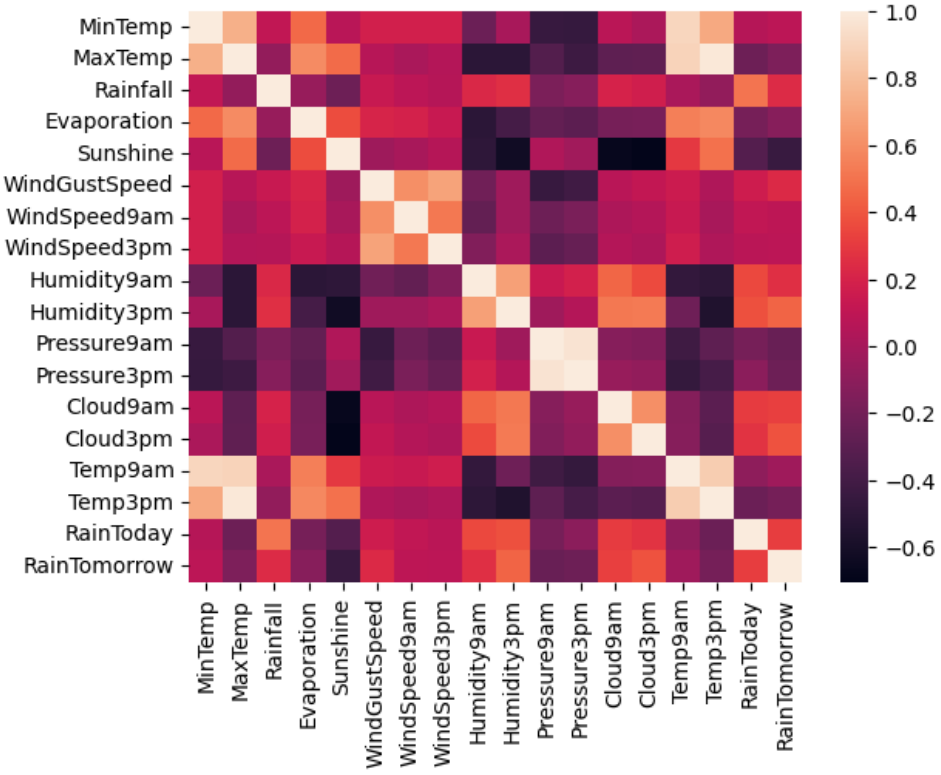
- WindSpeed3pm - uśredniona prędkość z 10 min przed 15 badanego dnia, mierzona w kilometrach na godzinę.
- Humidity9am - wilgotność o 9 rano badanego dnia, mierzona w procentach.
- Humidity3pm - wilgotność o 15 badanego dnia, mierzona w procentach.
- Pressure9am – presja atmosferyczna o 9 rano badanego dnia, zredukowana do średniego poziomu mórz, mierzona w hektopaskalach
- Pressure3pm – presja atmosferyczna o 15 badanego dnia, zredukowana do średniego poziomu mórz, mierzona w hektopaskalach
- Cloud9am - część nieba zakryta chmurami o 9 rano, mierzona w oktach.
- Cloud3pm - część nieba zakryta chmurami o 15, mierzona w oktach.
- RainToday – zmienna zero jedynkowa przyjmujące 1 gdy spadł deszcz w przeciągu 24 godzin od 9 rano dnia badanego, zero w przeciwnym przypadku.

Wyrzuciliśmy te zmienne przez to, że mają zbyt wiele unikalnych wartości, przez co powstałoby dużo zmiennych zero-jedynkowych:

- Date – data obserwacji
- Location – miejsce obserwacji
- WindGustDir – Kierunek najsilniejszego wiatru w ciągu 24 h od północy
- WindDir9am - Średni kierunek wiatru z 10 min przed 9 rano
- WindDir3pm - Średni kierunek wiatru z 10 min przed 15
- RISK_MM - ilość deszczu który spadł następnego dnia w milimetrach

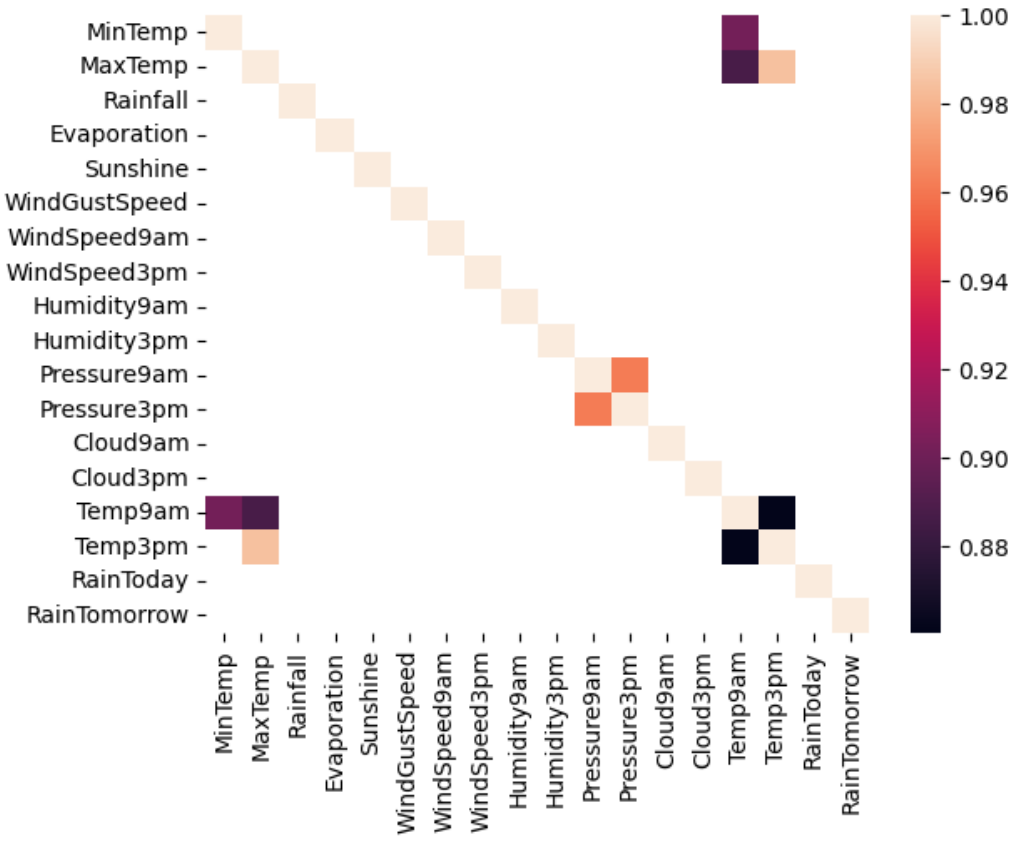
Następnie przeprowadzamy analizę korelacji pozostałych zmiennych:

	MinTemp	MaxTemp	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
MinTemp	1.000000	0.736267	0.467261	0.072961	0.177285	0.176005	0.175749	-0.234211	0.005999	-0.451260	-0.461623	0.077625	0.020489	0.901813	0.708865	0.056185	0.083936
MaxTemp	0.736267	1.000000	0.588915	0.469967	0.067690	0.014680	0.050800	-0.505432	-0.509270	-0.332293	-0.427279	-0.289865	-0.279053	0.887020	0.984562	-0.228884	-0.159237
Evaporation	0.467261	0.588915	1.000000	0.366607	0.203001	0.193936	0.128895	-0.505890	-0.392785	-0.269907	-0.293160	-0.185032	-0.184287	0.545497	0.574275	-0.187975	-0.119285
Sunshine	0.072961	0.469967	0.366607	1.000000	-0.032831	0.008040	0.056012	-0.491603	-0.629122	0.040959	-0.020464	-0.675610	-0.704202	0.291139	0.490180	-0.330635	-0.450768
WindGustSpeed	0.177285	0.067690	0.203001	-0.032831	1.000000	0.604837	0.686419	-0.215461	-0.026663	-0.457891	-0.412922	0.071235	0.109088	0.150258	0.032970	0.155490	0.234010
WindSpeed9am	0.176005	0.014680	0.193936	0.008040	0.604837	1.000000	0.519971	-0.270807	-0.031607	-0.227923	-0.174916	0.024280	0.053584	0.129298	0.005108	0.102267	0.090995
WindSpeed3pm	0.175749	0.050800	0.128895	0.056012	0.686419	0.519971	1.000000	-0.145942	0.015903	-0.295567	-0.254988	0.052780	0.025269	0.163601	0.028567	0.080074	0.087817
Humidity9am	-0.234211	-0.505432	-0.505890	-0.491603	-0.215461	-0.270807	-0.145942	1.000000	0.667388	0.139519	0.186955	0.452182	0.358043	-0.472826	-0.499777	0.353358	0.257161
Humidity3pm	0.005999	-0.509270	-0.392785	-0.629122	-0.026663	-0.031607	0.015903	0.667388	1.000000	-0.027449	0.051840	0.517037	0.523270	-0.221467	-0.557989	0.378766	0.446160
Pressure9am	-0.451260	-0.332293	-0.269907	0.040959	-0.457891	-0.227923	-0.295567	0.139519	-0.027449	1.000000	0.961348	-0.130081	-0.148139	-0.422773	-0.287301	-0.189804	-0.246371
Pressure3pm	-0.461623	-0.427279	-0.293160	-0.020464	-0.412922	-0.174916	-0.254988	0.186955	0.051840	0.961348	1.000000	-0.061152	-0.084963	-0.470325	-0.389863	-0.106298	-0.226031
Cloud9am	0.077625	-0.289865	-0.185032	-0.675610	0.071235	0.024280	0.052780	0.452182	0.517037	-0.130081	-0.061152	1.000000	0.604118	-0.137843	-0.302520	0.305950	0.317380
Cloud3pm	0.020489	-0.279053	-0.184287	-0.704202	0.109088	0.053584	0.025269	0.358043	0.523270	-0.148139	-0.084963	0.604118	1.000000	-0.127869	-0.318254	0.272149	0.381870
Temp9am	0.901813	0.887020	0.545497	0.291139	0.150258	0.129298	0.163601	-0.472826	-0.221467	-0.422773	-0.470325	-0.137843	-0.127869	1.000000	0.860574	-0.096593	-0.025691
Temp3pm	0.708865	0.984562	0.574275	0.490180	0.032970	0.005108	0.028567	-0.499777	-0.557989	-0.287301	-0.389863	-0.302520	-0.318254	0.860574	1.000000	-0.234925	-0.192424
RainToday	0.056185	-0.228884	-0.187975	-0.330635	0.155490	0.102267	0.080074	0.353358	0.378766	-0.189804	-0.106298	0.305950	0.272149	-0.096593	-0.234925	1.000000	0.313097
RainTomorrow	0.083936	-0.159237	-0.119285	-0.450768	0.234010	0.090995	0.087817	0.257161	0.446160	-0.246371	-0.226031	0.317380	0.381870	-0.025691	-0.192424	0.313097	1.000000



Jeśli weźmiemy pod uwagę tylko zmienne które są skorelowane z inną zmienną niż same z sobą w stopniu większym niż 80%.

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
MinTemp	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.901813	NaN	NaN	NaN
MaxTemp	NaN	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.887020	0.984562	NaN	NaN
Rainfall	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Evaporation	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Sunshine	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WindGustSpeed	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WindSpeed9am	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WindSpeed3pm	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Humidity9am	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Humidity3pm	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Pressure9am	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.961348	NaN	NaN	NaN	NaN	NaN	NaN
Pressure3pm	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.961348	1.000000	NaN	NaN	NaN	NaN	NaN	NaN
Cloud9am	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN
Cloud3pm	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN
Temp9am	0.901813	0.887020	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.860574	NaN	NaN
Temp3pm	NaN	0.984562	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.860574	1.000000	NaN	NaN
RainToday	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN
RainTomorrow	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0



Widzimy że należałoby wyeliminować zmienne:

- Temp9am – temperatura o 9 rano, mierzona w stopniach Celsjusza.
- Temp3pm – temperatura o 15, mierzona w stopniach Celsjusza.

Ze względu na to, że są ściśle związane z MinTemp oraz MaxTemp.

b) Statystyki opisowe

Wartości podstawowych statystyk czyli: średniej, odchylenia standardowego, minimum, maximum, mediany oraz kwartyli opisowych przedstawia poniższa tabelka:

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	RainToday	RainTomorrow
count	141556.000000	141871.000000	140787.000000	81350.000000	74377.000000	132923.000000	140845.000000	139563.000000	140419.000000	138583.000000	128179.000000	128212.000000	88536.000000	85099.000000	140787.000000	142193.000000
mean	12.186400	23.226784	2.349974	5.469824	7.624853	39.984292	14.001988	18.637576	68.843810	51.482606	1017.653758	1015.258204	4.437189	4.503167	0.223423	0.224181
std	6.403283	7.117618	8.465173	4.188537	3.781525	13.588801	8.893337	8.803345	19.051293	20.797772	7.105476	7.036677	2.887016	2.720633	0.416541	0.417043
min	-8.500000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000	0.000000	0.000000	980.500000	977.100000	0.000000	0.000000	0.000000	0.000000
25%	7.600000	17.900000	0.000000	2.600000	4.900000	31.000000	7.000000	13.000000	57.000000	37.000000	1012.900000	1010.400000	1.000000	2.000000	0.000000	0.000000
50%	12.000000	22.600000	0.000000	4.800000	8.500000	39.000000	13.000000	19.000000	70.000000	52.000000	1017.600000	1015.200000	5.000000	5.000000	0.000000	0.000000
75%	16.800000	28.200000	0.800000	7.400000	10.600000	48.000000	19.000000	24.000000	83.000000	66.000000	1022.400000	1020.000000	7.000000	7.000000	0.000000	0.000000
max	33.900000	48.100000	371.000000	145.000000	14.500000	135.000000	130.000000	87.000000	100.000000	100.000000	1041.000000	1039.600000	9.000000	9.000000	1.000000	1.000000

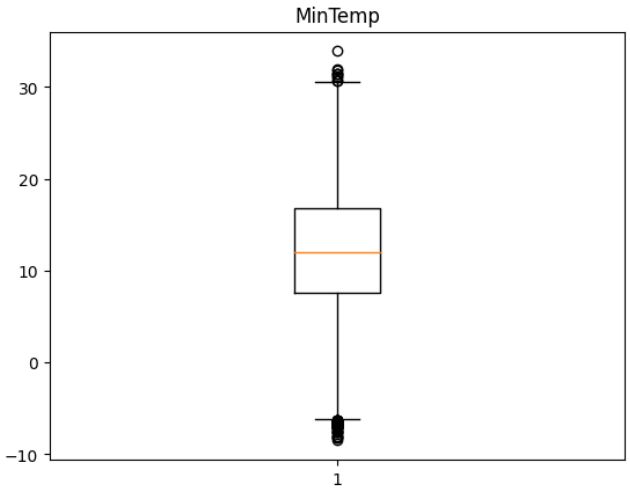
Dodatkowo patrząc na statystyki opisowe, uwagę przyciąga count wachający się od 74377 do 141871 jeśli bierzemy pod uwagę tylko zmienne objaśniające a 142193 jeśli weźmiemy też pod uwagę zmienną objaśnianą. Pokazuje to duże ilości braków danych w datasecie. Najwyższa ilość danych ze zmiennych

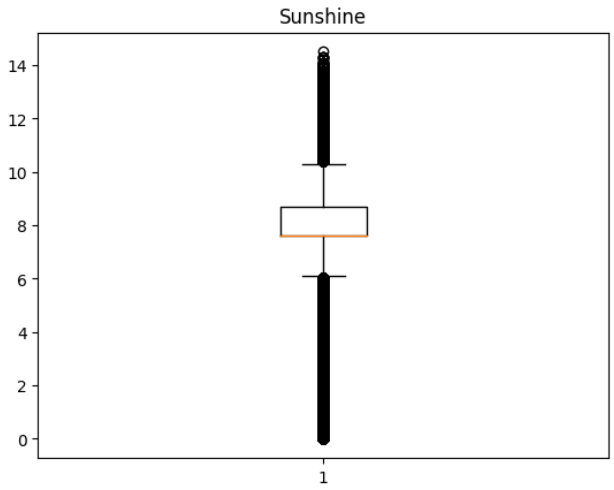
objaśniających zawiera zmienna RainToday a najmniejsza Sunshine. Wszystkie zmienne które nie zostały usunięte przez brak unikalności, są zmiennymi numerycznymi, dzięki czemu możemy zobaczyć że najniższą średnią ze zmiennych objaśniających posiada RainToday które jest zmienną zero jedynkową, a największą presure9am wynoszącą 1017.653758 hpa czyli nie wiele przekraczająca 1013 hPa czyli jedną atmosferę. Można tez zauważyć że wszystkie zmienne nie wyrażane w hektopaskalach i celsjuszach mają wartość minimalną równą zero.

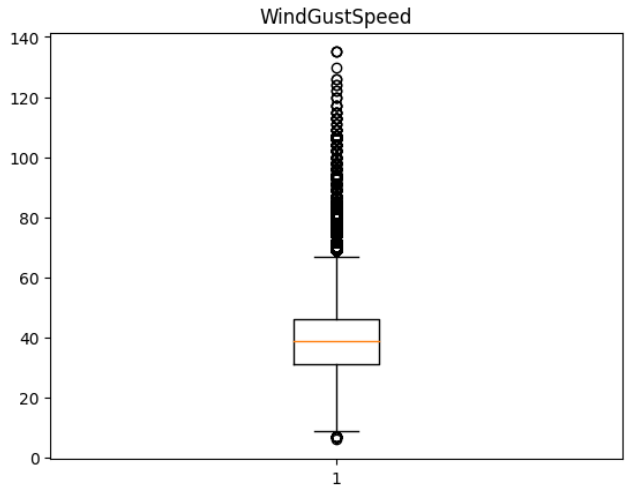
```
df.skew()

MinTemp          0.023900
MaxTemp          0.224917
Rainfall         9.888061
Evaporation      3.746834
Sunshine        -0.502911
WindGustSpeed    0.874305
WindSpeed9am     0.775494
WindSpeed3pm     0.631433
Humidity9am     -0.482821
Humidity3pm      0.034515
Pressure9am     -0.096211
Pressure3pm     -0.046198
Cloud9am        -0.224286
Cloud3pm        -0.224092
RainToday        1.327992
RainTomorrow     1.322753
dtype: float64
```

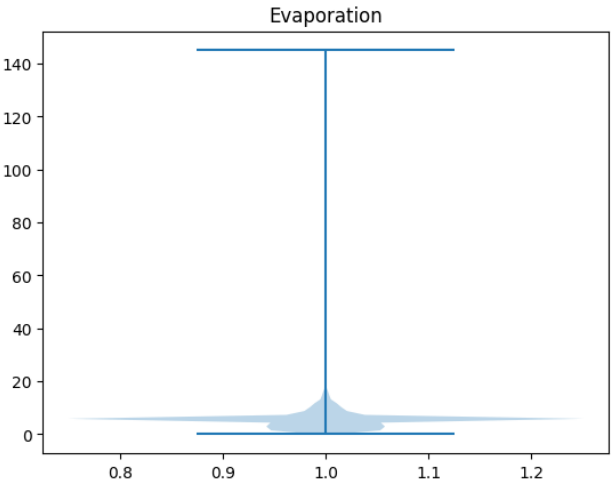
```
df.kurtosis()

MinTemp          -0.487253
MaxTemp          -0.238446
Rainfall        180.002097
Evaporation      45.067784
Sunshine         -0.820364
WindGustSpeed     1.417855
WindSpeed9am      1.226555
WindSpeed3pm      0.775865
Humidity9am      -0.039246
Humidity3pm      -0.511101
Pressure9am       0.236200
Pressure3pm       0.132521
Cloud9am         -1.541159
Cloud3pm         -1.457933
RainToday        -0.236441
RainTomorrow     -0.250329
dtype: float64
```

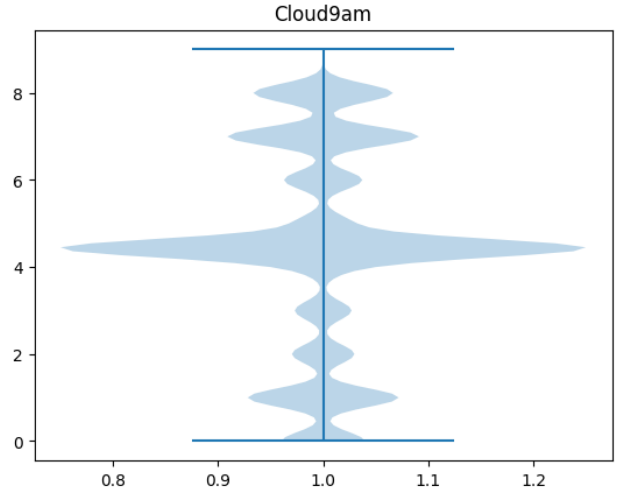
Jak widać zmienne na powyższym wycinku z programu MinTemp, MaxTemp, Sunshine, Humiidty9am i Humidity3pm, Cloud9am, Cloud3pm, Rain Today oraz Rain Tommorow mają ujemną kurtozę czyli ich rozkład jest platykurtyczny, przez co ich wartości mają wyższe rozproszenie od średniej niż w rozkładzie normalnym i dzięki temu prawdopodobieństwo wartości odstających jest mniejsze niż w rozkładzie normalnym. Reszta zmiennych ma dodatnią kurtozę przez co ich rozkład jest leptokurtyczny i mają większe prawdopodobieństwo występowania wartości odstających niż rozkład normalny. Zmiennę MinTemp, MaxTemp, Rainfall, Evaporation, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidoty3pm, Rain Today oraz Rain Tommorow mają dodatnią asymetrię czyli wykazuja się asymetrię prawostronną czyli większość obserwacji posiada wartość mniejsza niż średnia. Pozostałe zmienne mają ujemną asymetrię czyli lewostronną, oznacza to że większość obserwacji i mediana są większę od średniej.

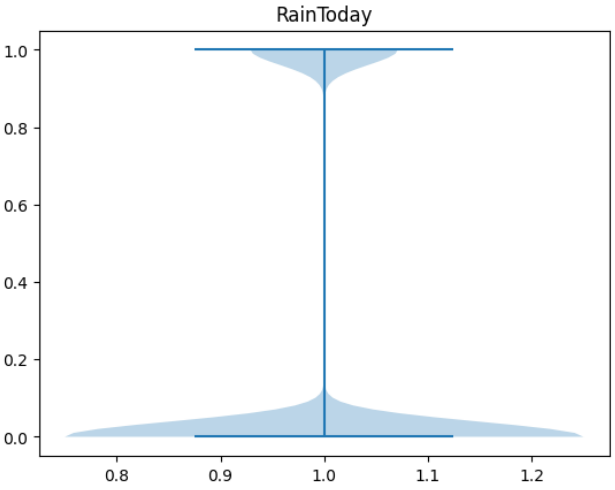

MinTemp
30
20
10
0
−10
1


Sunshine
14
12
10
8
6
4
2
0
1


WindGustSpeed
140
120
100
80
60
40
20
0
1

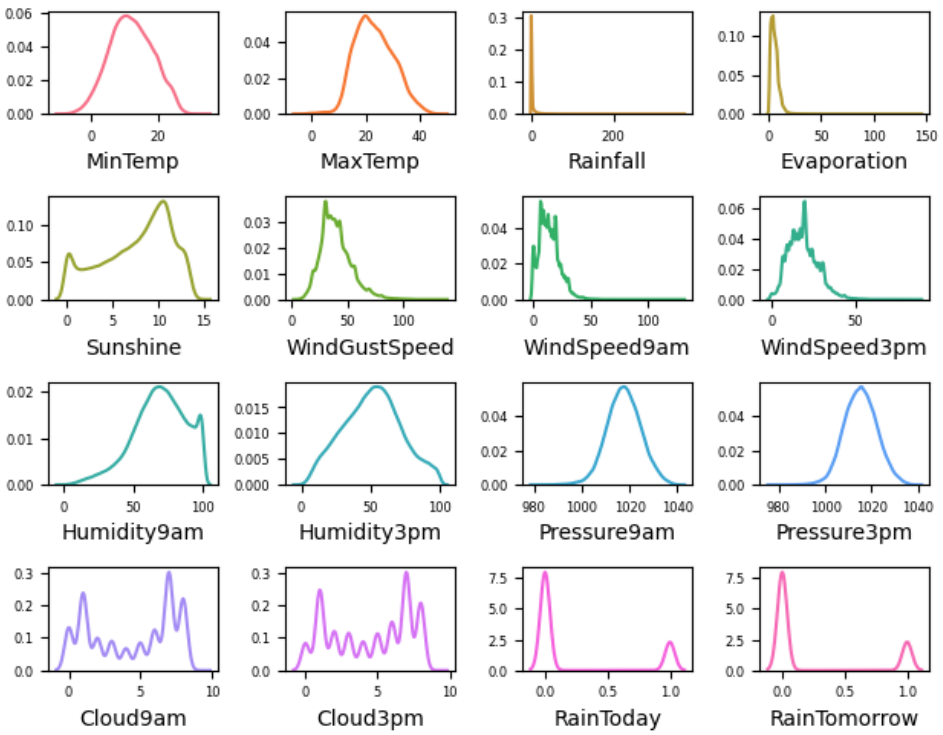

Evaporation
140
120
100
80
60
40
20
0
0.8
0.9
1.0
1.1
1.2

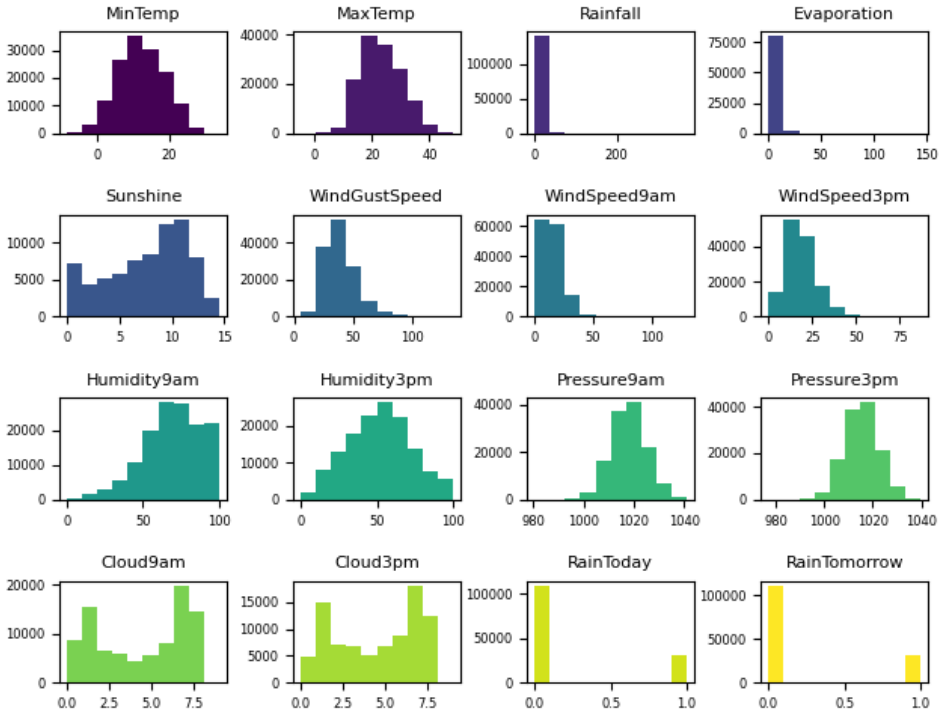

Cloud9am
8
6
4
2
0
0.8
0.9
1.0
1.1
1.2


RainToday
1.0
0.8
0.6
0.4
0.2
0.0
0.8
0.9
1.0
1.1
1.2

5.3 Wizualizacja

Rozkład zmiennych przed modyfikacjami można zobrazować zestawem wykresów dystrybucji jak i histogramów:





Dokonując analizy powyższych wykresów, nietrudno zauważyć, że zmienne mintemp i maxtemp charakteryzują się lekką asymetrią prawostronną (co dodatkowo potwierdza dodatnia acz bliska zeru wartość współczynnika skośności tych zmiennych). Po rozkładzach zmiennych evaporation i rainfall widać że są silnie asymetryczne prawostronnie, co również potwierdzają wysokie wartości współczynnika skośności. Wykresy dotyczące zmiennej sunshine pozwalają dojść do wniosku, że rozkład tej zmiennej jest lewostronnie asymetryczny i obserwacje zbliżone do minimum występują częściej niż w rozkładzie normalnym (można mówić o większym zbliżeniu do rozkładu dwumodalnego niż normalnego). Zmienne dotyczące prędkości wiatru (czyli WindGustSpeed, WindSpeed9am i WindSpeed3pm) charakteryzują się wręcz klasycznym przykładem asymetrii prawostronnej, przy czym najsilniej asymetryczna jest zmienna windspeeed9am. Ciekawym przypadkiem jest wilgotność, która rano (zmienna Humidity9Am) jest wyraźnie asymetryczna lewostronnie z lekkim wskazaniem na dwumodalność, zaś popołudniu (zmienna Humidity3pm) asymetria jest niezauważalna na wykresach. Wykresy opisujące zmienne dotyczące ciśnienia atmosferycznego (pressure9am i pressure3pm) wskazują na rozkład tych zmiennych zbliżony do normalnego. Zmienne dotyczące zachmurzenia (cloud 9am i cloud 3pm) są z rozkładu dwumodalnego na co wskazują po dwa wysokie słupki na histogramach niedaleko wartości minimalnych i maksymalnych. Zmienne raintoday i raintomorrow są zmiennymi binarnymi, w których większość obserwacji stanowią dni niedeszczowe.

5.4 Transformacje danych

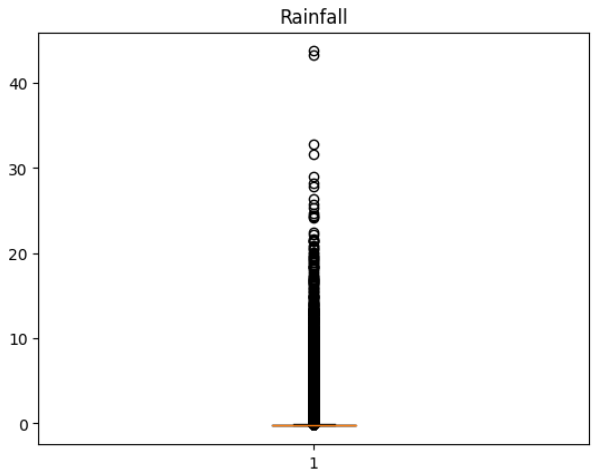
Wszystkie zmienne ciągłe zostały zestandaryzowane według klasycznego schematu, czyli odjęto od wartości empirycznej wartość średnią danej cechy, a następnie podzielono wartość przez odchylenie standardowe. Natomiast zmienna binarna RainToday pozostała niezmieniona.

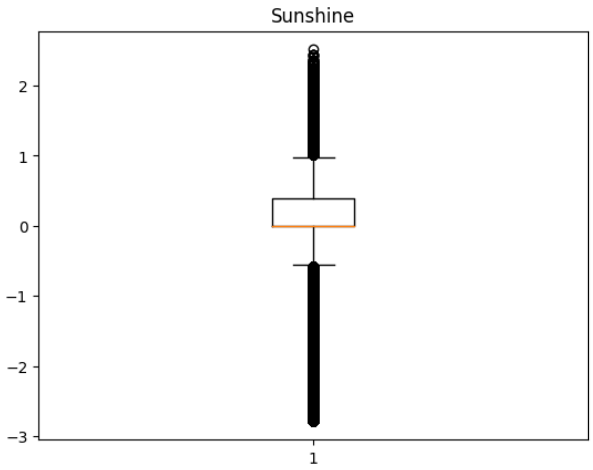
5.5 Obsługa braków danych

Braki danych w przypadku zmiennej binarnej RainToday zostały uzupełnione wartością najczęsciej występującą, ponieważ zastąpienie średnią nie dałoby wartości binarnej. W zmiennych ciągłych natomiast zastosowana została metoda uzupełniania wartością średnią każdej zmiennej.

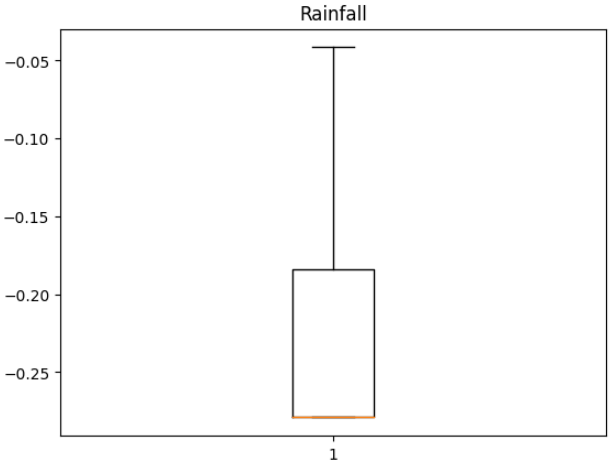
5.6 Obserwacje odstające

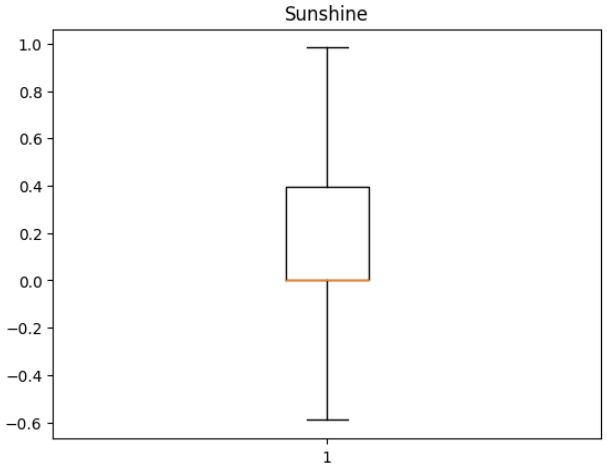
W naszym zbiorze wystąpiło sporo wartości odstających z powodu, iż nasz zbiór był ogromny, zatem postanowiliśmy nie usuwać wartości odstających. W tym celu wykorzystaliśmy boxploty aby zwizualizować problem.





Wartości odstające zastąpiliśmy odpowiednio dolnymi oraz górnymi wąsami w przypadku zmiennych ciągłych, a zmienną binarną RainToday pozostawiliśmy bez zmian. Zbiór zmiennych po transformacji prezentuje się następująco:



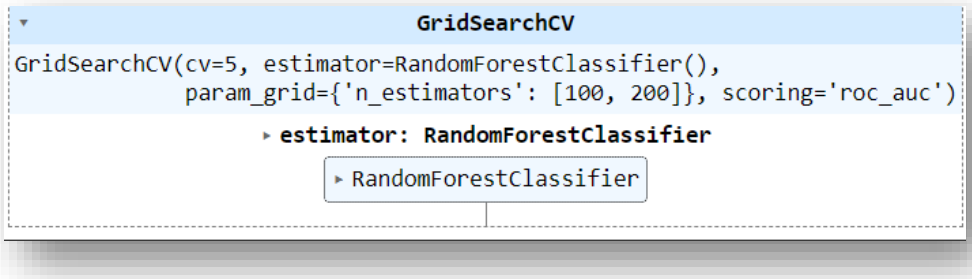


6. Opis metod

Do każdego z algorytmów zostały dobrane odpowiednie parametry za pomocą GridSearch'a, który sprawdza kombinacje parametrów za pomocą cross-walidacji na 5 podzbiorach i optymalizuje wartość metryki AUC. Następnie zapisywane są najlepsze parametry z podanego setu i przekazywane do klasyfikatorów. Dla każdego klasyfikatora zostały dobrane różne parametry oraz została zilustrowana krzywa ROC i macierz konfuzji.

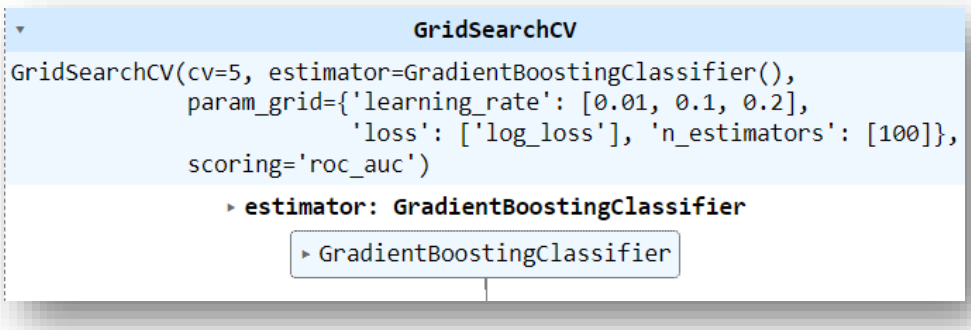
6.1 Lasy losowe

Lasy losowe polegają na uczeniu wielu drzew losowych na różnych podzbiorach zbioru treningowego a następnie łączy ich wyniki przez uśrednianie (Ensemble Voting). Metoda ta została zaproponowana przez Leo Breimana w pracy naukowej „Random Forests. Machine Learning" z 2001 roku i znalazła szerokie zastosowanie w badaniach naukowych z różnych dziedzin między innymi w pracy „PORÓWNANIE SKUTECZNOŚCI DWÓCH KULTUR ANALITYCZNYCH" autorstwa Bolesława Borkowskiego, Marka Karwańskiego i Wiesława Szczęsnego. Algorytm ten jest skuteczny dzięki losowemu wyborowi cech i losowemu podziałowi danych, co pomaga w zapobieganiu nadmiernemu dopasowaniu pojedynczych drzew(overfitting). Głosowanie odbywa się w sposób większościowy. Lasy losowe w przypadku naszego projektu nie były najszybszym algorytmem zważywszy na duży zbiór danych, więc ograniczyliśmy się do sprawdzania w GridSearch'u jedynie 2 wartości dla parametru n_estimators.



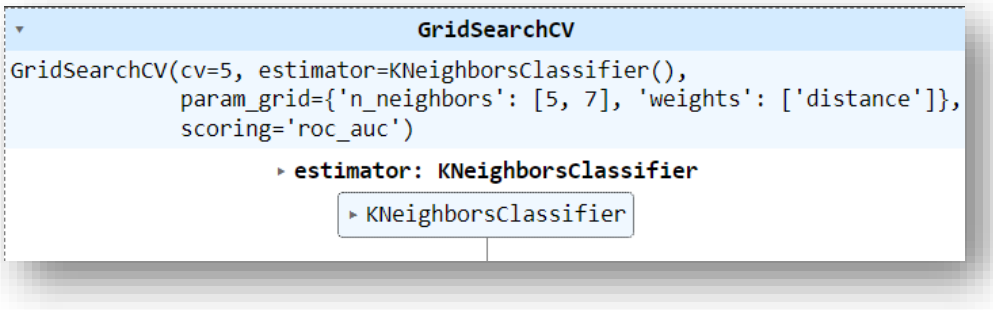
6.2 Wzmocnienie gradientowe (gradient boosting)

Gradient Boosting Classifier w bibliotece sklearn jest algorytmem zespołowym, który jak Random Forest działa na wielu drzewach losowych. W algorytmie tym minimalizujemy gradient funkcji straty. Ważnym elementem jest również iteracyjne dodawanie drzew, dostosowując wagę błędów biorąc pod uwagę poprzednie wyniki. W naszym wypadku tworzone jest 100 drzew losowych. Algorytm ten również podobnie jak algorytm lasów losowych został wykorzystany w pracy „PORÓWNANIE SKUTECZNOŚCI DWÓCH KULTUR ANALITYCZNYCH" napisanej przez Bolesława Borkowskiego, Marka Karwańskiego i Wiesława Szczęsnego.



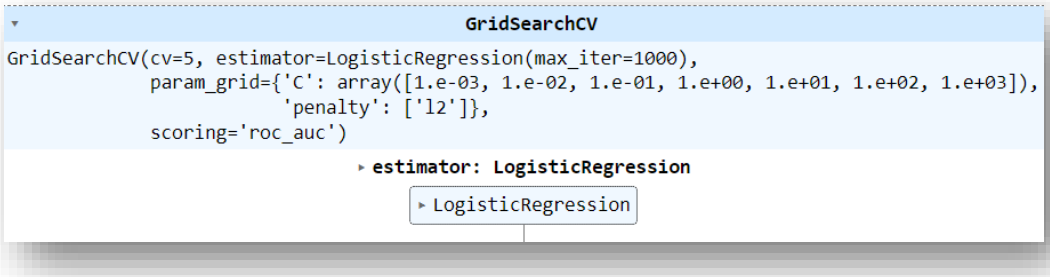
6.3 Metoda k najbliższych sąsiadów (KNN)

Algorytm KNN jest algorytmem, który przypisuje etykietę klasy dla nowej obserwacji biorąc pod uwagę jej otoczenie (K- najbliższych sąsiadów). Jest on szeroko stosowany w różnorakich pracach naukowych, czego przykładem jest „Lokalna ocena mocy dyskryminacyjnej zmiennych" autorstwa Mariusza Kubusa z uniwersytetu ekonomicznego we Wrocławiu. W przypadku naszego projektu algorytm KNN nie jest pożądany, ponieważ mamy bardzo duży zbiór danych, dlatego w GridSearch'u zdecydowaliśmy się przetestować jedynie 2 wartości parametru n_neighbors.



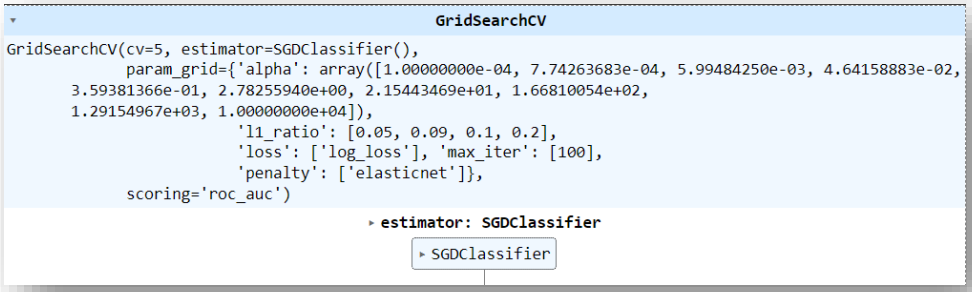
6.4 Regresja logistyczna

Regresją logistyczną nazywamy metodę używaną do modelowania zależności pomiędzy zmienną zależną a pozostałymi zmiennymi niezależnymi, gdzie zmienna zależna jest binarna. Celem użycia tej metody jest prawdopodobieństwo przewidzenia przynależności do jednej z dwóch różnych klas. Przykładem zastosowania regresji logistycznej do klasyfikacji w badaniach naukowych jest praca autorstwa Rafała Piszczka pod tytułem „Prognozowanie bankructwa z zastosowaniem modelu logitowego". W przypadku regresji logistycznej jako parametry GridSearch'a przetestowaliśmy kilka wartości dla zmiennej C, która jest parametrem w regularyzacji, która została przez nas wybrana jako l2.



6.5 Metoda stochastycznego spadku wzdłuż gradientu (SGD)

Ideą SGD jest aktualizacja parametrów danego modelu w oparciu o gradient funkcji straty (pochodne funkcji straty) na podstawie losowo wybranej próbki danych. Następnie dany proces jest powtarzany dla innych próbek do momentu, gdy osiągnięta zostanie zdefiniowana liczba iteracji bądź innych kryteriów zakończenia. Metoda została dokładnie opisana i przetestowana w pracy „Backpropagation and stochastic gradient descent method" autorstwa Prof. Shun-ichi Amari. W algorytmie SGD do GridSearch'a podaliśmy więcej parametrów, ponieważ jest to algorytm bardzo dobrze sprawdzający się na dużych zbiorach danych.



6.6 Głosowanie większościowe(twarde)

Głosowanie większościowe dla algorytmów klasyfikacji polega na klasyfikacji zmiennej poprzez wszystkie algorytmy samodzielnie, a następnie wybranie klasy, która była najczęściej prognozowana przez pojedyncze algorytmy. Algorytm głosowania większościowego został wykorzystany między innymi w pracy „Classifier selection for majority voting" napisanej przez Dymitra Ruta i Bogdana Gabrysa.

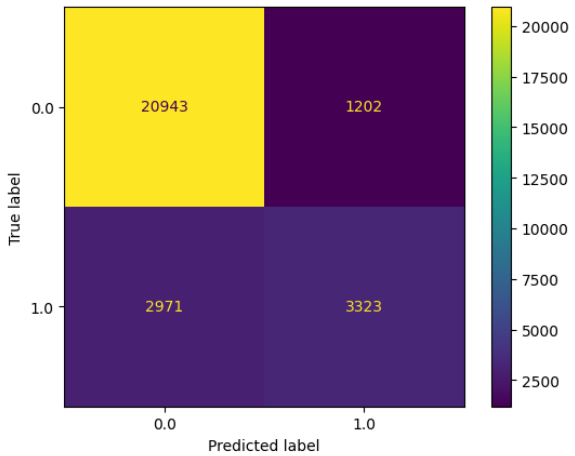
6.7 Głosowanie według najwyższego prawdopodobieństwa (głosowanie miękkie)

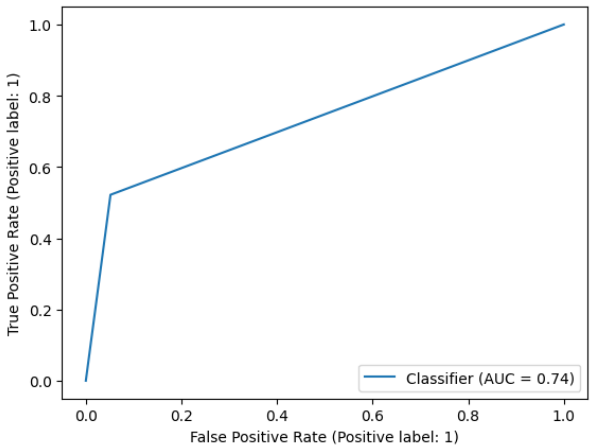
W głosowaniu miękkim w odróżnieniu do głosowania twardego algorytmy klasyfikujące nie zwracają etykiety klas, a prawdopodobieństwo przynależności zmiennej do klasy. W ten sposób klasa zmiennej jest ustalana na podstawie najwyższej wartości prawdopodobieństwa z modeli. Metoda głosowania miękkiego została wykorzystana w tekście „A new hybrid credit scoring ensemble model with feature enhancement and soft voting weight optimization" autorstwa Dongqi Yang**,** Binqing Xiao, Mengya Cao i**.** Huaqi Shen.

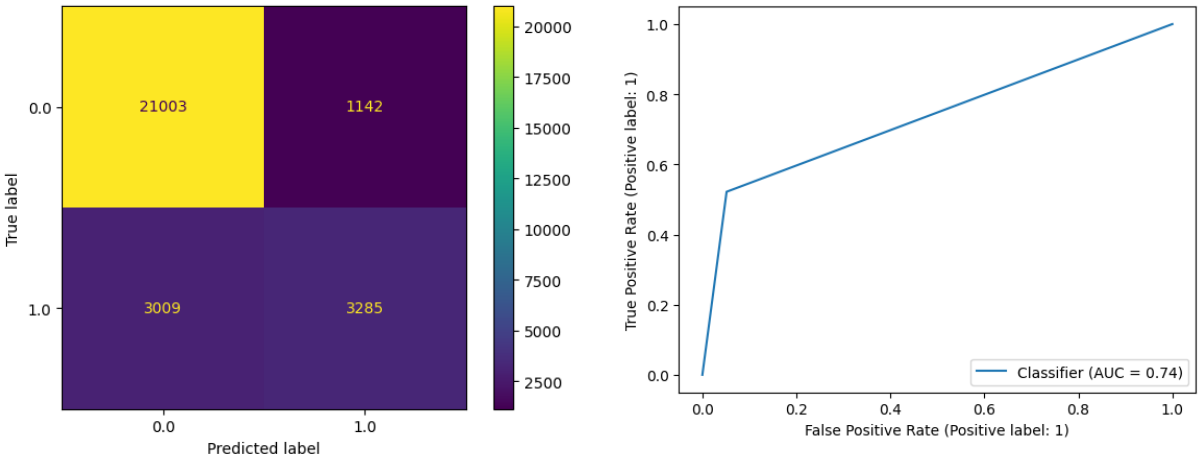
7. Rezultaty

7.1 Mierniki

W naszym projekcie postanowiliśmy użyć 2 mierników dokładności modeli. Pierwszym z nich oraz najpopularniejszym i najprostszym jest accuracy. Miernik ten jest wynikiem dzielenia liczby poprawnych predykcji przez wszystkie predykcje. Dzięki niemu można w prosty sposób wyciągnąć wnioski, wartość 1 oznacza idealne sklasyfikowanie każdego przypadku, natomiast wartość 0 oznacza brak jakiejkolwiek poprawnej etykiety zmiennej. Warto jednak dodać, że w przypadku, gdy w modelu jedna zmienna występuje znacznie częściej, accuracy może być wysokie ze względu na poprawną klasyfikację jednej cechy. Z tego powodu postanowiliśmy dodać również miernik AUC score. AUC- Area Under the Curve, czyli jest to pole pod krzywą ROC. Krzywa ta przedstawia stosunek czułości (Stosunek poprawnie sklasyfikowanej klasy 1 przez liczbę rzeczywistych obserwacji z klasy 1) do 1 minus specyficzności (Liczba wszystkich źle sklasyfikowanych obserwacji 1 klasy przez liczbę wszystkich złych predykcji) dla różnych progów decyzyjnych. Im bliżej krzywa ROC jest do lewego górnego rogu, tym lepszy jest model. Wartość bliska 1 oznacza idealne rozróżnianie klas przez algorytm, natomiast wartość bliska 0,5 oznacza losowe rozpoznawanie etykiety klasy. W naszym przypadku miernika accuracy wypadał dobrze lub bardzo dobrze szczególnie dla algorytmów opartych o głosowanie większościowe (SGD, GD, Random Forest). Natomiast w przypadku Australii znacznie częściej występują dni, w których nie ma opadów, około 5 razy częściej. Z tego powodu miernik AUC nie wypadł tak dobrze dla naszych algorytmów, jednak osiągał on wartość w granicy 0,75, co można uznać za akceptowalną zdolność do rozróżniania klas. Macierze konfuzji oraz krzywe ROC dla modeli hybrydowych prezentują się następująco:







Jak widać naszym model zdecydowanie lepiej radzi sobie z rozpoznawaniem dni bez deszczu, może być to wynikiem zdecydowanie mniejszej ilości dni deszczowych oraz złożonością samej pogody.

7.2 Sposób walidacji

Jako sposób walidacji wybraliśmy walidację krzyżową. Jest to metoda oceny wydajności modelu, który jest dzielony na kilka podzbiorów w naszym przypadku 5 i iteracyjnym trenowaniu oraz testowaniu na różnych kombinacjach tych podzbiorów. Metoda ta skutecznie pomaga uniknąć przetrenowania.

8. Przykład użycia modeli na stworzonych sztucznie obserwacjach

Stworzona została sztuczna zmienna o powyższych wartościach poszczególnych zmiennych i przetestowano na niej model hybrydowy drugi. Wynikiem predykcji była klasa 1 mówiąca o tym, że następnego dnia pojawią się opady deszczu.

```
Z = X.head(1)
Z['MinTemp'] = 20
Z['MaxTemp'] = 27
Z['Rainfall'] = 100
Z['Evaporation'] = 50
Z['Sunshine'] = 6
Z['WindGustSpeed'] = 20
Z['WindSpeed9am'] = 10
Z['WindSpeed3pm'] = 15
Z['Humidity9am'] = 70
Z['Humidity3pm'] = 60
Z['Pressure9am'] = 1008
Z['Pressure3pm'] = 1005
Z['Cloud9am'] = 7
Z['Cloud3pm'] = 7
Z = pd.DataFrame(scaler.transform(Z))
Z['RainToday'] = 1
Z.columns =  X_train.columns
Z
```

```
[251] model_clfs2.predict(Z)

      array([1.])
```

9. Bibliografia

Breiman L. (2001) Random Forests. Machine Learning, 45, 5-32

Perdał, R. (2018). ZASTOSOWANIE ANALIZY SKUPIEŃ I LASÓW LOSOWYCH W KLASYFIKACJI GMIN W POLSCE NA SKALI POZIOMU ROZWOJU SPOŁECZNO-GOSPODARCZEGO. *Metody Ilościowe W Badaniach Ekonomicznych*, *19*(3), 263–273. https://doi.org/10.22630/MIBE.2018.19.3.24

Borkowski, B., Karwański, M., & Szczesny, W. (2021). PORÓWNANIE SKUTECZNOŚCI DWÓCH KULTUR ANALITYCZNYCH. *Metody Ilościowe W Badaniach Ekonomicznych*, *22*(4), 124–138. https://doi.org/10.22630/MIBE.2021.22.4.11

Kubus M., 2016a, Lokalna ocena mocy dyskryminacyjnej zmiennych, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 427, Taksonomia 27, s. 143-152, DOI: 10.15611/pn.2016.427.15.

Piszczek, R., (2012) Prognozowanie bankructwa z zastosowaniem modelu logitowego, Studia Ekonomiczne / Uniwersytet Ekonomiczny w Katowicach nr 94 Perspektywy rozwoju gospodarki regionalnej : analizy ekonometryczno-statystyczne s 117-131.

Shun-ichi Amari (1993) „Backpropagation and stochastic gradient descent method", Neurocomputing Volume 5, Issues 4–5, June 1993, s185-196

Dymitr Ruta, Bogdan Gabrys, „Classifier selection for majority voting", Information Fusion Volume 6, Issue 1, March 2005, Pages 63-81

Dongqi Yang**,** Binqing Xiao, Mengya Cao i**.** Huaqi Shen „A new hybrid credit scoring ensemble model with feature enhancement and soft voting weight optimization" Expert Systems with Applications Volume 238, Part C, 15 March 2024, 122101