In [191]:
```
#import seaborn as sns
import csv
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.model_selection import StratifiedShuffleSplit,train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
```

In [2]:
```
def print_full(x):
    pd.set_option('display.max_rows', len(x))
    pd.set_option('display.max_columns', 500)
    pd.set_option('display.width', 2000)
    pd.set_option('display.float_format', '{:20,.2f}'.format)
    pd.set_option('display.max_colwidth', -1)
    print(x)
    pd.reset_option('display.max_rows')
    pd.reset_option('display.max_columns')
    pd.reset_option('display.width')
    pd.reset_option('display.float_format')
    pd.reset_option('display.max_colwidth')
```

In [3]:
```
data = pd.read_csv("weatherAUS.csv")
data.head(10)
```

| | Date | Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustDir | WindGustSpee |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | NaN | NaN | W | 44.0 |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | NaN | NaN | WNW | 44.0 |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | NaN | NaN | WSW | 46.0 |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NaN | NaN | NE | 24.0 |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | NaN | NaN | W | 41.0 |
| 5 | 2008-12-06 | Albury | 14.6 | 29.7 | 0.2 | NaN | NaN | WNW | 56.0 |
| 6 | 2008-12-07 | Albury | 14.3 | 25.0 | 0.0 | NaN | NaN | W | 50.0 |
| 7 | 2008-12-08 | Albury | 7.7 | 26.7 | 0.0 | NaN | NaN | W | 35.0 |
| 8 | 2008-12-09 | Albury | 9.7 | 31.9 | 0.0 | NaN | NaN | NNW | 80.0 |
| 9 | 2008-12-10 | Albury | 13.1 | 30.1 | 1.4 | NaN | NaN | W | 28.0 |

10 rows × 24 columns

In [4]:
```
1 data
```

| | Date | Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustDir | WindGust |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | NaN | NaN | W | 44.0 |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | NaN | NaN | WNW | 44.0 |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | NaN | NaN | WSW | 46.0 |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NaN | NaN | NE | 24.0 |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | NaN | NaN | W | 41.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142188 | 2017-06-20 | Uluru | 3.5 | 21.8 | 0.0 | NaN | NaN | E | 31.0 |
| 142189 | 2017-06-21 | Uluru | 2.8 | 23.4 | 0.0 | NaN | NaN | E | 31.0 |
| 142190 | 2017-06-22 | Uluru | 3.6 | 25.3 | 0.0 | NaN | NaN | NNW | 22.0 |
| 142191 | 2017-06-23 | Uluru | 5.4 | 26.9 | 0.0 | NaN | NaN | N | 37.0 |
| 142192 | 2017-06-24 | Uluru | 7.8 | 27.0 | 0.0 | NaN | NaN | SE | 28.0 |

142193 rows × 24 columns

In [5]:
```
1 Braki = data.isna().sum()  # suma brakujacych wartosci w kazdej z kolumn
```

In [6]:
```
1 Braki/142193 * 100   #procent  NaN
```

```
Date             0.000000
Location         0.000000
MinTemp          0.447983
MaxTemp          0.226453
Rainfall         0.988797
Evaporation     42.789026
Sunshine        47.692924
WindGustDir      6.561504
WindGustSpeed    6.519308
WindDir9am       7.041838
WindDir3pm       2.656952
WindSpeed9am     0.948007
WindSpeed3pm     1.849599
Humidity9am      1.247600
Humidity3pm      2.538803
Pressure9am      9.855619
Pressure3pm      9.832411
Cloud9am        37.735332
Cloud3pm        40.152469
Temp9am          0.635756
Temp3pm          1.917113
RainToday        0.988797
RISK_MM          0.000000
RainTomorrow     0.000000
dtype: float64
```

In [7]:
```
1 data = data.drop( labels=["Cloud3pm","Cloud9am","Sunshine","Evaporation","RISK_M
2                           axis=1, inplace=False, errors='raise') #riskmm = rainfall,
```

In [8]:
```
1 data #usuniecie kolumn
```

| | Date | Location | MinTemp | MaxTemp | Rainfall | WindGustDir | WindGustSpeed | WindDir9am | Wi |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | W | 44.0 | W | WN |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | WNW | 44.0 | NNW | WS |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | WSW | 46.0 | W | WS |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NE | 24.0 | SE | E |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | W | 41.0 | ENE | NW |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142188 | 2017-06-20 | Uluru | 3.5 | 21.8 | 0.0 | E | 31.0 | ESE | E |
| 142189 | 2017-06-21 | Uluru | 2.8 | 23.4 | 0.0 | E | 31.0 | SE | ENI |
| 142190 | 2017-06-22 | Uluru | 3.6 | 25.3 | 0.0 | NNW | 22.0 | SE | N |
| 142191 | 2017-06-23 | Uluru | 5.4 | 26.9 | 0.0 | N | 37.0 | SE | WN |
| 142192 | 2017-06-24 | Uluru | 7.8 | 27.0 | 0.0 | SE | 28.0 | SSE | N |

142193 rows × 19 columns

In [9]:
```
1 data.head(50)
```

| | Date | Location | MinTemp | MaxTemp | Rainfall | WindGustDir | WindGustSpeed | WindDir9am | WindDi |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | W | 44.0 | W | WNW |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | WNW | 44.0 | NNW | WSW |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | WSW | 46.0 | W | WSW |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NE | 24.0 | SE | E |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | W | 41.0 | ENE | NW |
| 5 | 2008-12-06 | Albury | 14.6 | 29.7 | 0.2 | WNW | 56.0 | W | W |
| 6 | 2008-12-07 | Albury | 14.3 | 25.0 | 0.0 | W | 50.0 | SW | W |
| 7 | 2008-12-08 | Albury | 7.7 | 26.7 | 0.0 | W | 35.0 | SSE | W |
| 8 | 2008-12-09 | Albury | 9.7 | 31.9 | 0.0 | NNW | 80.0 | SE | NW |
| 9 | 2008-12-10 | Albury | 13.1 | 30.1 | 1.4 | W | 28.0 | S | SSE |
| 10 | 2008-12-11 | Albury | 13.4 | 30.4 | 0.0 | N | 30.0 | SSE | ESE |
| 11 | 2008-12-12 | Albury | 15.9 | 21.7 | 2.2 | NNE | 31.0 | NE | ENE |
| 12 | 2008-12-13 | Albury | 15.9 | 18.6 | 15.6 | W | 61.0 | NNW | NNW |
| 13 | 2008-12-14 | Albury | 12.6 | 21.0 | 3.6 | SW | 44.0 | W | SSW |
| 14 | 2008-12-16 | Albury | 9.8 | 27.7 | NaN | WNW | 50.0 | NaN | WNW |
| 15 | 2008-12-17 | Albury | 14.1 | 20.9 | 0.0 | ENE | 22.0 | SSW | E |
| 16 | 2008-12-18 | Albury | 13.5 | 22.9 | 16.8 | W | 63.0 | N | WNW |
| 17 | 2008-12-19 | Albury | 11.2 | 22.5 | 10.6 | SSE | 43.0 | WSW | SW |
| 18 | 2008-12-20 | Albury | 9.8 | 25.6 | 0.0 | SSE | 26.0 | SE | NNW |
| 19 | 2008-12-21 | Albury | 11.5 | 29.3 | 0.0 | S | 24.0 | SE | SE |
| 20 | 2008-12-22 | Albury | 17.1 | 33.0 | 0.0 | NE | 43.0 | NE | N |
| 21 | 2008-12-23 | Albury | 20.5 | 31.8 | 0.0 | WNW | 41.0 | W | W |
| 22 | 2008-12-24 | Albury | 15.3 | 30.9 | 0.0 | N | 33.0 | ESE | NW |
| 23 | 2008-12-25 | Albury | 12.6 | 32.4 | 0.0 | W | 43.0 | E | W |
| 24 | 2008-12-26 | Albury | 16.2 | 33.9 | 0.0 | WSW | 35.0 | SE | WSW |
| 25 | 2008-12-27 | Albury | 16.9 | 33.0 | 0.0 | WSW | 57.0 | NaN | W |
| 26 | 2008-12-28 | Albury | 20.1 | 32.7 | 0.0 | WNW | 48.0 | N | WNW |
| 27 | 2008-12-29 | Albury | 19.7 | 27.2 | 0.0 | WNW | 46.0 | NW | WSW |
| 28 | 2008-12-30 | Albury | 12.5 | 24.2 | 1.2 | WNW | 50.0 | WSW | SW |
| 29 | 2008-12-31 | Albury | 12.0 | 24.4 | 0.8 | W | 39.0 | WNW | WNW |
| 30 | 2009-01-01 | Albury | 11.3 | 26.5 | 0.0 | WNW | 56.0 | W | WNW |
| 31 | 2009-01-02 | Albury | 9.6 | 23.9 | 0.0 | W | 41.0 | WSW | SSW |
| 32 | 2009-01-03 | Albury | 10.5 | 28.8 | 0.0 | SSE | 26.0 | SSE | E |
| 33 | 2009-01-04 | Albury | 12.3 | 34.6 | 0.0 | WNW | 37.0 | SSE | NW |
| 34 | 2009-01-05 | Albury | 12.9 | 35.8 | 0.0 | WNW | 41.0 | ENE | NW |
| 35 | 2009-01-06 | Albury | 13.7 | 37.9 | 0.0 | W | 52.0 | SE | WNW |
| 36 | 2009-01-07 | Albury | 16.1 | 38.9 | 0.0 | W | 57.0 | E | W |
| 37 | 2009-01-08 | Albury | 14.0 | 28.3 | 0.0 | W | 48.0 | W | WSW |
| 38 | 2009-01-09 | Albury | 12.5 | 28.4 | 0.0 | NE | 37.0 | SSE | S |
| 39 | 2009-01-10 | Albury | 17.0 | 30.8 | 0.0 | NE | 37.0 | NNE | E |
| 40 | 2009-01-11 | Albury | 16.9 | 32.0 | 0.0 | S | 31.0 | SSE | N |

In [10]:
```
Braki = data.isna().sum()
Braki
```

```
Date                0
Location            0
MinTemp           637
MaxTemp           322
Rainfall         1406
WindGustDir      9330
WindGustSpeed    9270
WindDir9am      10013
WindDir3pm       3778
WindSpeed9am     1348
WindSpeed3pm     2630
Humidity9am      1774
Humidity3pm      3610
Pressure9am     14014
Pressure3pm     13981
Temp9am           904
Temp3pm          2726
RainToday        1406
RainTomorrow        0
dtype: int64
```

In [ ]:
```

```

# Imputancja

In [11]:
```
MinTemp_median          = data['MinTemp'].median()      #wyznaczenie miedian i dom
MaxTemp_median          = data['MaxTemp'].median()
Rainfall_median         = data['Rainfall'].median()
WindGustDir_dominant    = data['WindGustDir'].mode()
WindGustSpeed_median    = data['WindGustSpeed'].median()
WindDir9am_dominant     = data['WindDir9am'].mode()
WindDir3pm_dominant     = data['WindDir3pm'].mode()
WindSpeed9am_median     = data['WindSpeed9am'].median()
WindSpeed3pm_median     = data['WindSpeed3pm'].median()
Humidity9am_median      = data['Humidity9am'].median()
Humidity3pm_median      = data['Humidity3pm'].median()
Pressure9am_median      = data['Pressure9am'].median()
Pressure3pm_median      = data['Pressure3pm'].median()
Temp9am_median          = data['Temp9am'].median()
Temp3pm_median          = data['Temp3pm'].median()
RainToday_dominant      =  data['RainToday'].mode()
```

In [12]:
```
WindGustDir_dominant.to_numpy()[0]
```

```
'W'
```

In [13]:
```
WindGustSpeed_median
```

```
39.0
```

In [14]:
```
data["MinTemp"].fillna(MinTemp_median, inplace = True) # uzupełnienie wartości
data['MaxTemp'].fillna(MaxTemp_median, inplace = True)
data['Rainfall'].fillna(Rainfall_median, inplace = True)
data['WindGustDir'].fillna(WindGustDir_dominant.to_numpy()[0], inplace = True)
data['WindGustSpeed'].fillna(WindGustSpeed_median, inplace = True)
data['WindDir9am'].fillna(WindDir9am_dominant.to_numpy()[0], inplace = True)
data['WindDir3pm'].fillna(WindDir3pm_dominant.to_numpy()[0], inplace = True)
data['WindSpeed9am'].fillna(WindSpeed9am_median, inplace = True)
data['WindSpeed3pm'].fillna(WindSpeed3pm_median, inplace = True)
data['Humidity9am'].fillna(Humidity9am_median, inplace = True)
data['Humidity3pm'].fillna(Humidity3pm_median, inplace = True)
data['Pressure9am'].fillna(Pressure9am_median, inplace = True)
data['Pressure3pm'].fillna(Pressure3pm_median, inplace = True)
data['Temp9am'].fillna(Temp9am_median, inplace = True)
data['Temp3pm'].fillna(Temp3pm_median, inplace = True)
data['RainToday'].fillna(RainToday_dominant.to_numpy()[0], inplace = True)
```

In [15]:
```
print_full(data['WindGustDir'][3460:3470])
```

```
3460    NNE
3461    E
3462    WSW
3463    E
3464    ENE
3465    W
3466    ESE
3467    NNE
3468    NE
3469    SSW
Name: WindGustDir, dtype: object
```

In [16]:
```
1 Braki = data.isna().sum()
2 Braki
```

```
Date             0
Location         0
MinTemp          0
MaxTemp          0
Rainfall         0
WindGustDir      0
WindGustSpeed    0
WindDir9am       0
WindDir3pm       0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Temp9am          0
Temp3pm          0
RainToday        0
RainTomorrow     0
dtype: int64
```

In [17]:

```
def mod_outlier(df):                                    #liczenie IQR
      df1 = df.copy()
      df = df._get_numeric_data()


      q1 = df.quantile(0.25)
      q3 = df.quantile(0.75)

      iqr = q3 - q1
      display(iqr)

      lower_bound = q1 -(1.5 * iqr)
      upper_bound = q3 +(1.5 * iqr)


      for col in df.columns:
          for i in range(0,len(df[col])):
              if df[col][i] < lower_bound[col]:
                  df[col][i] = 0.0000000000000000001

              if df[col][i] > upper_bound[col]:
                  df[col][i] = 0.0000000000000000001


      for col in df.columns:
          df1[col] = df[col]

      return(df1)
```

In [18]:

```
odcieta_data = mod_outlier(data)
```

```
MinTemp          9.2
MaxTemp         10.3
Rainfall         0.6
WindGustSpeed   15.0
WindSpeed9am    12.0
WindSpeed3pm    11.0
Humidity9am     26.0
Humidity3pm     28.0
Pressure9am      8.3
Pressure3pm      8.4
Temp9am          9.2
Temp3pm          9.6
dtype: float64
```

In [19]:

```
odcieta_data
```

| | Date | Location | MinTemp | MaxTemp | Rainfall | WindGustDir | WindGustSpeed | WindDir9am | Wi |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | W | 44.0 | W | WN |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | WNW | 44.0 | NNW | WS |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | WSW | 46.0 | W | WS |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NE | 24.0 | SE | E |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | W | 41.0 | ENE | NW |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142188 | 2017-06-20 | Uluru | 3.5 | 21.8 | 0.0 | E | 31.0 | ESE | E |
| 142189 | 2017-06-21 | Uluru | 2.8 | 23.4 | 0.0 | E | 31.0 | SE | ENI |
| 142190 | 2017-06-22 | Uluru | 3.6 | 25.3 | 0.0 | NNW | 22.0 | SE | N |
| 142191 | 2017-06-23 | Uluru | 5.4 | 26.9 | 0.0 | N | 37.0 | SE | WN |
| 142192 | 2017-06-24 | Uluru | 7.8 | 27.0 | 0.0 | SE | 28.0 | SSE | N |

142193 rows × 19 columns

In [20]:

```
odcieta_data= odcieta_data[odcieta_data.MinTemp != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.MaxTemp  != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Rainfall != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.WindGustSpeed != 0.00000000000000000001
odcieta_data = odcieta_data[odcieta_data.WindSpeed9am != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.WindSpeed3pm != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Humidity9am != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Humidity3pm  != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Pressure9am != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Pressure3pm != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Temp9am != 0.0000000000000000001 ]
odcieta_data = odcieta_data[odcieta_data.Temp3pm!= 0.0000000000000000001 ]
```

In [21]:
```
display(odcieta_data)  #df bez tych wystajacych nad quantile
```

| | Date | Location | MinTemp | MaxTemp | Rainfall | WindGustDir | WindGustSpeed | WindDir9am | Wir |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | W | 44.0 | W | WN |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | WNW | 44.0 | NNW | WS |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | WSW | 46.0 | W | WS |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NE | 24.0 | SE | E |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | W | 41.0 | ENE | NW |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142188 | 2017-06-20 | Uluru | 3.5 | 21.8 | 0.0 | E | 31.0 | ESE | E |
| 142189 | 2017-06-21 | Uluru | 2.8 | 23.4 | 0.0 | E | 31.0 | SE | ENE |
| 142190 | 2017-06-22 | Uluru | 3.6 | 25.3 | 0.0 | NNW | 22.0 | SE | N |
| 142191 | 2017-06-23 | Uluru | 5.4 | 26.9 | 0.0 | N | 37.0 | SE | WN |
| 142192 | 2017-06-24 | Uluru | 7.8 | 27.0 | 0.0 | SE | 28.0 | SSE | N |

105507 rows × 19 columns

In [22]:
```
kategorie = odcieta_data.loc[:,['Location',"WindGustDir",'WindDir9am','WindDir3p
kategorie
```

| | Location | WindGustDir | WindDir9am | WindDir3pm | RainToday |
|---|---|---|---|---|---|
| 0 | Albury | W | W | WNW | No |
| 1 | Albury | WNW | NNW | WSW | No |
| 2 | Albury | WSW | W | WSW | No |
| 3 | Albury | NE | SE | E | No |
| 4 | Albury | W | ENE | NW | No |
| ... | ... | ... | ... | ... | ... |
| 142188 | Uluru | E | ESE | E | No |
| 142189 | Uluru | E | SE | ENE | No |
| 142190 | Uluru | NNW | SE | N | No |
| 142191 | Uluru | N | SE | WNW | No |
| 142192 | Uluru | SE | SSE | N | No |

105507 rows × 5 columns

In [23]:
```
odcieta_data['Year']=[d.split('-')[0] for d in odcieta_data.Date]
odcieta_data['Month']=[d.split('-')[1] for d in odcieta_data.Date]
odcieta_data['Day']=[d.split('-')[2] for d in odcieta_data.Date]
```

In [24]:
```
odcieta_data = odcieta_data.drop( labels='Date',
                      axis=1, inplace=False, errors='raise')
odcieta_data  # edycja daty rozdzielenie na year month day
```

| | Location | MinTemp | MaxTemp | Rainfall | WindGustDir | WindGustSpeed | WindDir9am | WindDir3pm |
|---|---|---|---|---|---|---|---|---|
| 0 | Albury | 13.4 | 22.9 | 0.6 | W | 44.0 | W | WNW |
| 1 | Albury | 7.4 | 25.1 | 0.0 | WNW | 44.0 | NNW | WSW |
| 2 | Albury | 12.9 | 25.7 | 0.0 | WSW | 46.0 | W | WSW |
| 3 | Albury | 9.2 | 28.0 | 0.0 | NE | 24.0 | SE | E |
| 4 | Albury | 17.5 | 32.3 | 1.0 | W | 41.0 | ENE | NW |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142188 | Uluru | 3.5 | 21.8 | 0.0 | E | 31.0 | ESE | E |
| 142189 | Uluru | 2.8 | 23.4 | 0.0 | E | 31.0 | SE | ENE |
| 142190 | Uluru | 3.6 | 25.3 | 0.0 | NNW | 22.0 | SE | N |
| 142191 | Uluru | 5.4 | 26.9 | 0.0 | N | 37.0 | SE | WNW |
| 142192 | Uluru | 7.8 | 27.0 | 0.0 | SE | 28.0 | SSE | N |

105507 rows × 21 columns

In [25]:
```
odcieta_data
```

| | Location | MinTemp | MaxTemp | Rainfall | WindGustDir | WindGustSpeed | WindDir9am | WindDir3pm |
|---|---|---|---|---|---|---|---|---|
| 0 | Albury | 13.4 | 22.9 | 0.6 | W | 44.0 | W | WNW |
| 1 | Albury | 7.4 | 25.1 | 0.0 | WNW | 44.0 | NNW | WSW |
| 2 | Albury | 12.9 | 25.7 | 0.0 | WSW | 46.0 | W | WSW |
| 3 | Albury | 9.2 | 28.0 | 0.0 | NE | 24.0 | SE | E |
| 4 | Albury | 17.5 | 32.3 | 1.0 | W | 41.0 | ENE | NW |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142188 | Uluru | 3.5 | 21.8 | 0.0 | E | 31.0 | ESE | E |
| 142189 | Uluru | 2.8 | 23.4 | 0.0 | E | 31.0 | SE | ENE |
| 142190 | Uluru | 3.6 | 25.3 | 0.0 | NNW | 22.0 | SE | N |
| 142191 | Uluru | 5.4 | 26.9 | 0.0 | N | 37.0 | SE | WNW |
| 142192 | Uluru | 7.8 | 27.0 | 0.0 | SE | 28.0 | SSE | N |

105507 rows × 21 columns

In [26]:
```
x = odcieta_data.loc[:,['MinTemp',"MaxTemp",'Rainfall','WindGustSpeed','WindSpee
                                'WindSpeed3pm','Humidity9am','Humidity3pm','Press
                        'Pressure3pm','Temp9am','Temp3pm','Year','Month','Day']]

min_max_scaler = preprocessing.MinMaxScaler()
x_scaled = min_max_scaler.fit_transform(x)
normalized_data = pd.DataFrame(x_scaled, columns=x.columns)
```

In [ ]:
```
1
```

In [27]:
```
1 normalized_data #znormalizowana data numeric
```

| | MinTemp | MaxTemp | Rainfall | WindGustSpeed | WindSpeed9am | WindSpeed3pm | Humidity9am | Hur |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.538462 | 0.496333 | 0.400000 | 0.603448 | 0.540541 | 0.615385 | 0.646341 | 0.21 |
| 1 | 0.373626 | 0.550122 | 0.000000 | 0.603448 | 0.108108 | 0.564103 | 0.317073 | 0.24 |
| 2 | 0.524725 | 0.564792 | 0.000000 | 0.637931 | 0.513514 | 0.666667 | 0.243902 | 0.29 |
| 3 | 0.423077 | 0.621027 | 0.000000 | 0.258621 | 0.297297 | 0.230769 | 0.329268 | 0.15 |
| 4 | 0.651099 | 0.726161 | 0.666667 | 0.551724 | 0.189189 | 0.512821 | 0.780488 | 0.32 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 105502 | 0.266484 | 0.469438 | 0.000000 | 0.379310 | 0.405405 | 0.333333 | 0.500000 | 0.26 |
| 105503 | 0.247253 | 0.508557 | 0.000000 | 0.379310 | 0.351351 | 0.282051 | 0.402439 | 0.23 |
| 105504 | 0.269231 | 0.555012 | 0.000000 | 0.224138 | 0.351351 | 0.230769 | 0.463415 | 0.20 |
| 105505 | 0.318681 | 0.594132 | 0.000000 | 0.482759 | 0.243243 | 0.230769 | 0.426829 | 0.23 |
| 105506 | 0.384615 | 0.596577 | 0.000000 | 0.327586 | 0.351351 | 0.179487 | 0.402439 | 0.23 |

105507 rows × 15 columns

In [28]:
```
1 enc = preprocessing.OneHotEncoder()
2 enc.fit(kategorie)
3 onehotlabels = enc.transform(kategorie).toarray()
4 onehotlabels.shape
```

```
(105507, 99)
```

In [29]:
```
1 zakodowane= pd.DataFrame(onehotlabels)
2 zakodowane
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 |
| 1 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 |
| 2 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 |
| 3 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 4 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 105502 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 105503 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 105504 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 105505 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 |
| 105506 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

105507 rows × 99 columns

In [30]:

```
abc = pd.DataFrame(odcieta_data.loc[:,'Location'].values, columns = ['Location']
bca  = pd.DataFrame(odcieta_data.loc[:,'RainTomorrow'].values, columns = ['RainT
Big_boi = normalized_data.join(zakodowane)
Big_boi = abc.join(Big_boi)
Big_boi = bca.join(Big_boi)
Big_boi
```

| | RainTomorrow | Location | MinTemp | MaxTemp | Rainfall | WindGustSpeed | WindSpeed9am | WindSpe |
|---|---|---|---|---|---|---|---|---|
| 0 | No | Albury | 0.538462 | 0.496333 | 0.400000 | 0.603448 | 0.540541 | 0.615385 |
| 1 | No | Albury | 0.373626 | 0.550122 | 0.000000 | 0.603448 | 0.108108 | 0.564103 |
| 2 | No | Albury | 0.524725 | 0.564792 | 0.000000 | 0.637931 | 0.513514 | 0.666667 |
| 3 | No | Albury | 0.423077 | 0.621027 | 0.000000 | 0.258621 | 0.297297 | 0.230769 |
| 4 | No | Albury | 0.651099 | 0.726161 | 0.666667 | 0.551724 | 0.189189 | 0.512821 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 105502 | No | Uluru | 0.266484 | 0.469438 | 0.000000 | 0.379310 | 0.405405 | 0.333333 |
| 105503 | No | Uluru | 0.247253 | 0.508557 | 0.000000 | 0.379310 | 0.351351 | 0.282051 |
| 105504 | No | Uluru | 0.269231 | 0.555012 | 0.000000 | 0.224138 | 0.351351 | 0.230769 |
| 105505 | No | Uluru | 0.318681 | 0.594132 | 0.000000 | 0.482759 | 0.243243 | 0.230769 |
| 105506 | No | Uluru | 0.384615 | 0.596577 | 0.000000 | 0.327586 | 0.351351 | 0.179487 |

105507 rows × 116 columns

In [31]:

```
Big_boi = Big_boi.replace(list(set(Big_boi['RainTomorrow'])),\
                          [True, False]).astype({'RainTomorrow': 'bool'})
Big_boi["RainTomorrow"] = Big_boi["RainTomorrow"].astype(int)
Big_boi
```

| | RainTomorrow | Location | MinTemp | MaxTemp | Rainfall | WindGustSpeed | WindSpeed9am | WindSpe |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Albury | 0.538462 | 0.496333 | 0.400000 | 0.603448 | 0.540541 | 0.615385 |
| 1 | 1 | Albury | 0.373626 | 0.550122 | 0.000000 | 0.603448 | 0.108108 | 0.564103 |
| 2 | 1 | Albury | 0.524725 | 0.564792 | 0.000000 | 0.637931 | 0.513514 | 0.666667 |
| 3 | 1 | Albury | 0.423077 | 0.621027 | 0.000000 | 0.258621 | 0.297297 | 0.230769 |
| 4 | 1 | Albury | 0.651099 | 0.726161 | 0.666667 | 0.551724 | 0.189189 | 0.512821 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 105502 | 1 | Uluru | 0.266484 | 0.469438 | 0.000000 | 0.379310 | 0.405405 | 0.333333 |
| 105503 | 1 | Uluru | 0.247253 | 0.508557 | 0.000000 | 0.379310 | 0.351351 | 0.282051 |
| 105504 | 1 | Uluru | 0.269231 | 0.555012 | 0.000000 | 0.224138 | 0.351351 | 0.230769 |
| 105505 | 1 | Uluru | 0.318681 | 0.594132 | 0.000000 | 0.482759 | 0.243243 | 0.230769 |
| 105506 | 1 | Uluru | 0.384615 | 0.596577 | 0.000000 | 0.327586 | 0.351351 | 0.179487 |

105507 rows × 116 columns

In [32]:
```
def Regionowanko(df):
    df = df.copy()
    Regiony = np.unique(abc.values)
    lista_df = []
    for Region in Regiony:
        a = df.loc[df["Location"] == Region]
        lista_df.append(a)
    return lista_df
```

In [33]:
```
lista_region_df = Regionowanko(Big_boi)
```

In [83]:
```
len(lista_region_df)
```

```
49
```

In [34]:
```
lista_region_df[0]
```

| | RainTomorrow | Location | MinTemp | MaxTemp | Rainfall | WindGustSpeed | WindSpeed9am | WindSpee |
|---|---|---|---|---|---|---|---|---|
| 69267 | 1 | Adelaide | 0.519231 | 0.322738 | 0.533333 | 0.448276 | 0.351351 | 0.384615 |
| 69268 | 1 | Adelaide | 0.340659 | 0.305623 | 0.000000 | 0.189655 | 0.054054 | 0.282051 |
| 69269 | 1 | Adelaide | 0.315934 | 0.325183 | 0.000000 | 0.362069 | 0.162162 | 0.333333 |
| 69270 | 0 | Adelaide | 0.480769 | 0.320293 | 0.000000 | 0.741379 | 0.405405 | 0.564103 |
| 69271 | 0 | Adelaide | 0.428571 | 0.347188 | 0.000000 | 0.775862 | 0.459459 | 0.717949 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 71585 | 1 | Adelaide | 0.304945 | 0.403423 | 0.000000 | 0.172414 | 0.054054 | 0.230769 |
| 71586 | 1 | Adelaide | 0.293956 | 0.405868 | 0.000000 | 0.137931 | 0.000000 | 0.282051 |
| 71587 | 1 | Adelaide | 0.296703 | 0.413203 | 0.000000 | 0.224138 | 0.000000 | 0.179487 |
| 71588 | 1 | Adelaide | 0.318681 | 0.388753 | 0.000000 | 0.172414 | 0.108108 | 0.102564 |
| 71589 | 1 | Adelaide | 0.307692 | 0.359413 | 0.000000 | 0.137931 | 0.000000 | 0.230769 |

2323 rows × 116 columns

In [181]:

```
def startyfikacja(lista_region_df):
    lista_modeli = []
    lista_x_test = []
    lista_y_test = []
    lista_accu_region = []
    for region in lista_region_df:
        stratified = StratifiedShuffleSplit(n_splits = 4, test_size=200, random_

        df1 = region.copy()
        Rain = df1['RainTomorrow']
        df1.drop(labels=["RainTomorrow",'Location'],axis=1, inplace=True, errors

        for train_index, test_index in stratified.split(df1,Rain):
            x_train = df1.iloc[train_index]
            y_train = Rain.iloc[train_index]
            x_test =  df1.iloc[test_index]
            y_test = Rain.iloc[test_index]

            log = LogisticRegression(random_state=0,solver='lbfgs')
            log.max_iter = 1000
            log.fit(x_train, y_train)

            print(log.n_iter_)
            print(log.score(x_test,y_test))

            lista_modeli.append(log)
            lista_x_test.append(x_test)
            lista_y_test.append(y_test)
            lista_accu_region.append(log.score(x_test,y_test))

    return lista_modeli,lista_x_test,lista_y_test,lista_accu_region
```

In [182]:
```
modele,x_test,deszcz_test,lista_accu = startyfikacja(lista_region_df)
```

```
[81]
0.85
[82]
0.905
[75]
0.905
[84]
0.9
[85]
0.84
[107]
0.815
[101]
0.78
[86]
0.825
[80]
0.89
[85]
0.895
[74]
0.89
[90]
```

In [ ]:
```

```

In [183]:
```
x_test[4]
```

| | MinTemp | MaxTemp | Rainfall | WindGustSpeed | WindSpeed9am | WindSpeed3pm | Humidity9am | Hum |
|---|---|---|---|---|---|---|---|---|
| 80391 | 0.587912 | 0.396088 | 0.666667 | 0.517241 | 0.054054 | 0.487179 | 0.926829 | 0.515 |
| 79170 | 0.612637 | 0.449878 | 0.000000 | 0.517241 | 0.054054 | 0.000000 | 0.560976 | 0.666 |
| 79154 | 0.500000 | 0.374083 | 0.533333 | 0.517241 | 0.351351 | 0.487179 | 0.634146 | 0.515 |
| 78378 | 0.541209 | 0.452323 | 0.933333 | 0.517241 | 0.108108 | 0.487179 | 0.804878 | 0.616 |
| 80182 | 0.500000 | 0.342298 | 0.000000 | 0.517241 | 0.351351 | 0.487179 | 0.634146 | 0.515 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 79229 | 0.417582 | 0.303178 | 0.533333 | 0.517241 | 0.351351 | 1.000000 | 0.817073 | 0.727 |
| 78476 | 0.370879 | 0.288509 | 0.800000 | 0.517241 | 0.054054 | 0.102564 | 0.560976 | 0.606 |
| 80121 | 0.538462 | 0.469438 | 0.000000 | 0.517241 | 0.108108 | 0.487179 | 0.743902 | 0.515 |
| 80287 | 0.609890 | 0.535452 | 0.000000 | 0.517241 | 0.702703 | 0.487179 | 0.451220 | 0.515 |
| 80295 | 0.500000 | 0.523227 | 0.000000 | 0.517241 | 0.459459 | 0.487179 | 0.573171 | 0.515 |

200 rows × 114 columns

In [184]:
```
len(modele)
```

```
196
```

In [200]:

```
def sprawdz(modele,x_test,deszcz_test):
    test_size = 200
    XTest = pd.DataFrame()
    YTest = pd.DataFrame()
    lista_wynikow = []

    for x in x_test:
        XTest = XTest.append(x.head(test_size))


    for y in deszcz_test:
        YTest = YTest.append(pd.DataFrame(y).head(test_size))


    for i in modele:
        lista_wynikow.append(i.score(XTest,YTest))
    return lista_wynikow,XTest,YTest
```

In [201]:

```
Wyniki,X_test_glob,Y_test_glob = sprawdz(modele,x_test,deszcz_test)      #Wynikow
Wyniki = pd.DataFrame(Wyniki)
Wyniki.sort_values(0)
```

| | 0 |
|---|---|
| 125 | 0.799311 |
| 127 | 0.800102 |
| 119 | 0.800918 |
| 118 | 0.801786 |
| 126 | 0.802730 |
| ... | ... |
| 56 | 0.856122 |
| 106 | 0.856352 |
| 57 | 0.856837 |
| 59 | 0.857066 |
| 105 | 0.857092 |

196 rows × 1 columns

In [187]:
```
Wyniki_region = pd.DataFrame(lista_accu)
Wyniki_region.sort_values(0)
```

| | 0 |
|---|---|
| 109 | 0.745 |
| 6 | 0.780 |
| 110 | 0.790 |
| 135 | 0.800 |
| 74 | 0.800 |
| ... | ... |
| 14 | 0.960 |
| 164 | 0.960 |
| 194 | 0.965 |
| 192 | 0.970 |
| 193 | 0.970 |

196 rows × 1 columns

In [188]:
```
index = Wyniki.idxmax() #najwyższe acc na globalnym zbiorze
index_region = Wyniki_region.idxmax() #najwyzsze acc na zbiorze z regionu
display(index)
display(index_region)
```

```
0    105
dtype: int64
```

```
0    192
dtype: int64
```

In [ ]:
```
1
```

Wyzsza skutecznosc zostala osiagnieta na zbiorze z regionu, dlatego ze na tym był trenowany, o najwyzszej skutecznosci w calym kraju, nie jest najlepszy regionalnie, co jest rozsądne, bo by dostosowany do regionu, chyba, że byłby to bardzo stabilny pogodowo region.

In [ ]:
```
1
```

In [203]:
```
confusion_matrix(Y_test_glob, modele[105].predict(X_test_glob)) #confusion matri
#najlepszego globalnie klasyfikatora
```

```
array([[ 1153,  4891],
       [  711, 32445]], dtype=int64)
```

Skutecznosc na wysokosci 85%, jest dość satysfakcjonująca jak na statyczność danych dostarc klasyfikatora, posiadajac taki zbiór danych jesteśmy w stanie dość swobodnie "zgadywać" wysta dniu następnym, statystycznie w roku zmokniemy tylko 54 razy :).

In [ ]:
```
1
```