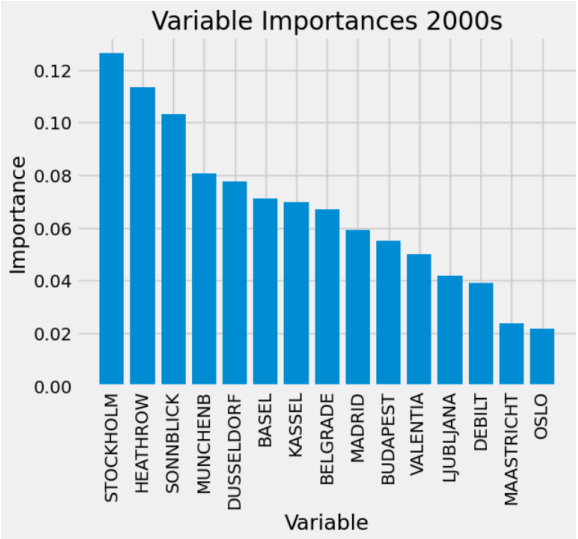
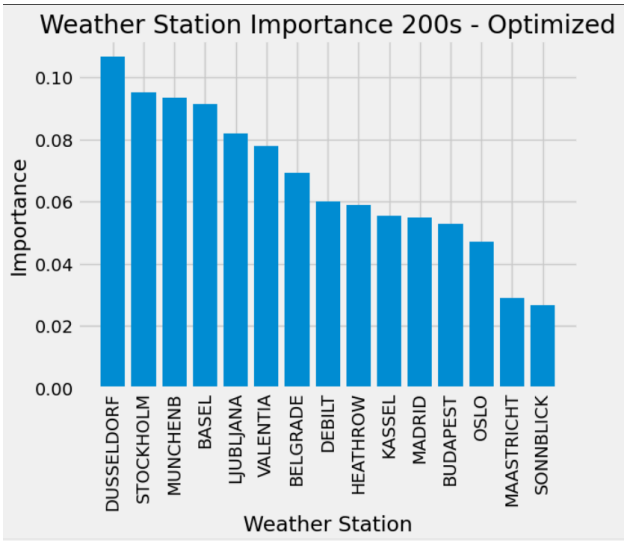
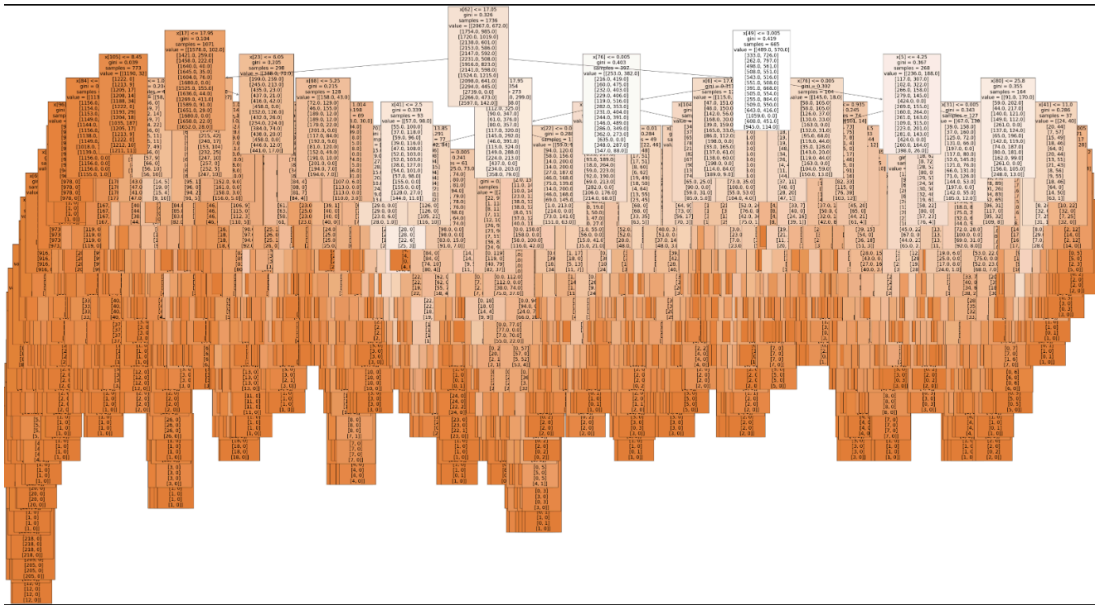
Exercise 2.4: Evaluating Hyperparameters

RandomForest

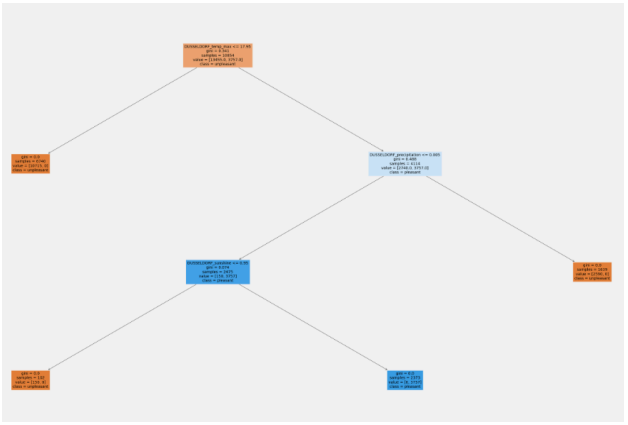
Exercise 2.3	Optimized
	

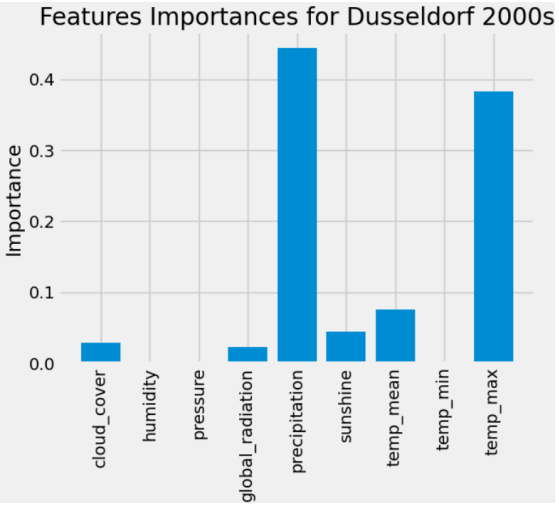
There is a noticeable variation in the importances of each station. Stockholm and Munchenb are the two only important variables that stay on top of the list after the optimization, but still, all stations change positions in the ranking.



The accuracy of the Optimized Random Forest is of 58%, just a bit higher that the 55% accuracy from the previous exercise.

Focusing on the most important station for the optimized model, Dusseldorf, the RandomForest results in a perfect accuracy of 100%. Considering that precipitation and maximum temperature take up almos 90% importance, it makes sense that there is such a high accuracy because there is not so much confusion on how other parameters affect to whether a day is pleasant or not. Still 100% is too perfect and this could indicate issues with the model.





CNN

Optimized

```
180/180 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
Pred        BASEL  BELGRADE  BUDAPEST  DEBILT  DUSSELDORF  HEATHROW  KASSEL  \
True
BASEL        3556        73        12      10           5         5       0
BELGRADE       75       983        12       4           0         3       0
BUDAPEST       22        16       155       6           1         4       0
DEBILT          7         4         9      58           2         2       0
DUSSELDORF      1         0         1       2          12         7       0
HEATHROW       12         1         0       1           2        36       0
KASSEL          1         1         1       0           2         0       3
LJUBLJANA       6         1         3       0           0         0       0
MAASTRICHT      6         0         0       0           0         0       0
MADRID         41         3         9       0           0         1       0
MUNCHENB        7         0         0       0           0         0       0
OSLO            0         0         0       0           0         0       0
STOCKHOLM       2         0         0       0           0         0       0
VALENTIA        1         0         0       0           0         0       0

Pred        LJUBLJANA  MAASTRICHT  MADRID  MUNCHENB  OSLO
True
BASEL               0           0      21         0     0
BELGRADE            4           0      11         0     0
BUDAPEST            1           0       9         0     0
DEBILT              0           0       0         0     0
DUSSELDORF          1           0       5         0     0
HEATHROW            1           0      27         0     2
KASSEL              0           1       1         1     0
LJUBLJANA          37           0      13         0     1
MAASTRICHT          0           2       1         0     0
MADRID              2           0     402         0     0
MUNCHENB            0           0       0         1     0
OSLO                0           0       0         0     5
STOCKHOLM           0           0       0         1     1
VALENTIA            0           0       0         0     0
```

Exercise 2.2

```
180/180 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step
Pred        BASEL  BELGRADE  BUDAPEST  DEBILT  DUSSELDORF  HEATHROW  KASSEL  \
True
BASEL           2        68       864      78         248        61     200
BELGRADE        0        89        77       4          78         1       6
BUDAPEST        0        15         9       3          21         0       0
DEBILT          0         2         0       0          14         0       0
DUSSELDORF      0         0         1       0           5         0       0
HEATHROW        0         6         2       3          13         0       1
KASSEL          0         1         0       0           1         0       0
LJUBLJANA       0         4         0       1           8         0       0
MAASTRICHT      0         0         0       2           2         0       1
MADRID          0        28        50      10          35         4       4
MUNCHENB        0         1         0       1           1         0       0
OSLO            0         2         0       0           0         0       0
STOCKHOLM       0         2         0       0           1         0       0
VALENTIA        0         0         0       0           0         0       0

Pred        LJUBLJANA  MAASTRICHT  MADRID  MUNCHENB  OSLO  SONNBLICK  \
True
BASEL            1183           1     303       413   115          6
BELGRADE          642           0      20       142     9          0
BUDAPEST          120           0       9        18     3          0
DEBILT             49           0       2        11     0          0
DUSSELDORF         16           0       1         4     2          0
HEATHROW           36           0       8         9     1          0
KASSEL              7           0       0         0     1          0
LJUBLJANA          34           0       8         1     2          0
MAASTRICHT          2           0       0         2     0          0
MADRID            134           0     139        28    11          0
MUNCHENB            3           0       0         1     1          0
OSLO                2           0       0         1     0          0
STOCKHOLM           0           0       0         0     0          0
VALENTIA            0           0       0         0     0          0

Pred        STOCKHOLM  VALENTIA
True
BASEL             103        37
BELGRADE           24         0
BUDAPEST           16         0
DEBILT              4         0
DUSSELDORF          0         0
HEATHROW            3         0
KASSEL              1         0
LJUBLJANA           3         0
MAASTRICHT          0         0
```

The optimized model doesn't recognize all 15 stations but does improve accuracy drastically, at 91.5%.

Iteration

If I were to segment data into smaller components I would first focus in creating regions out of the weather stations for this analysis. This would help categorize pleasant and unpleasant days in a much more accurate way considering this specific region's normal climate. Another way I could consider breaking down data would be classifying the days for season, a pleasant day in winter does not look the same as a pleasant day in summer.

As for what model to choose, both models have their pros and cons. An optimized random forest is more easily interpretable, and less prone to overfitting, but it struggles with complex patterns. On the other hand, optimized CNN is great at more complex, non-linear relationships, but it's harder to interpret.

I would suggest starting with the optimized random forest for weather predictions, especially for accuracy at individual stations.

Key variables:

- Precipitation
- Maximum temperature
- Mean temperature