Exercise 2.4: Evaluating Hyperparameters

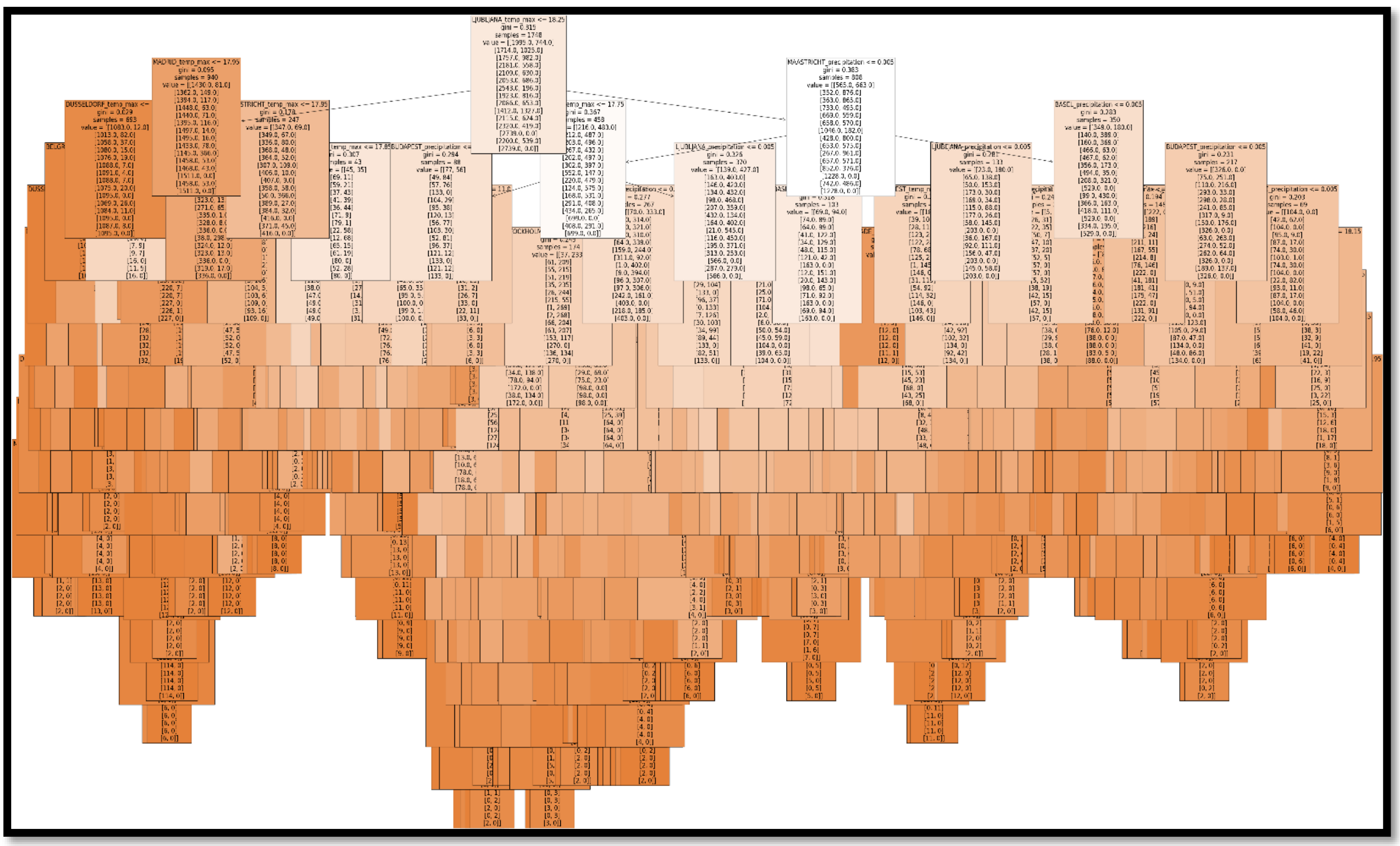
11 December 2024

Innocent Bayai

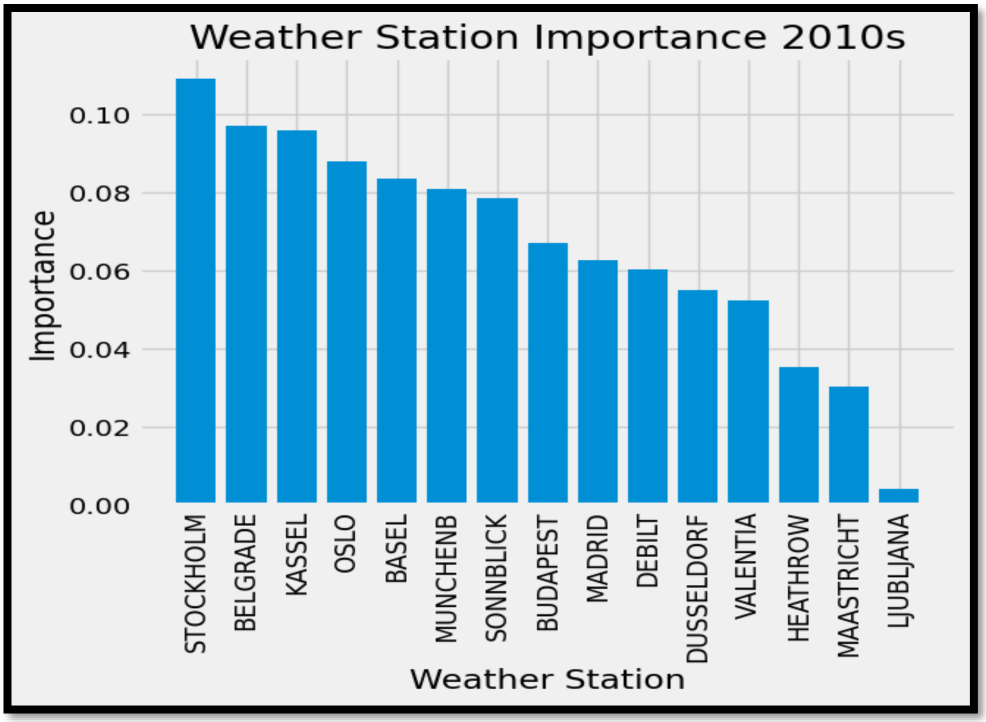
Part 1

The optimized random forest recorded an accuracy of approximately 66%, higher than the 63.89% recorded prior optimization through grid and randomized methods.

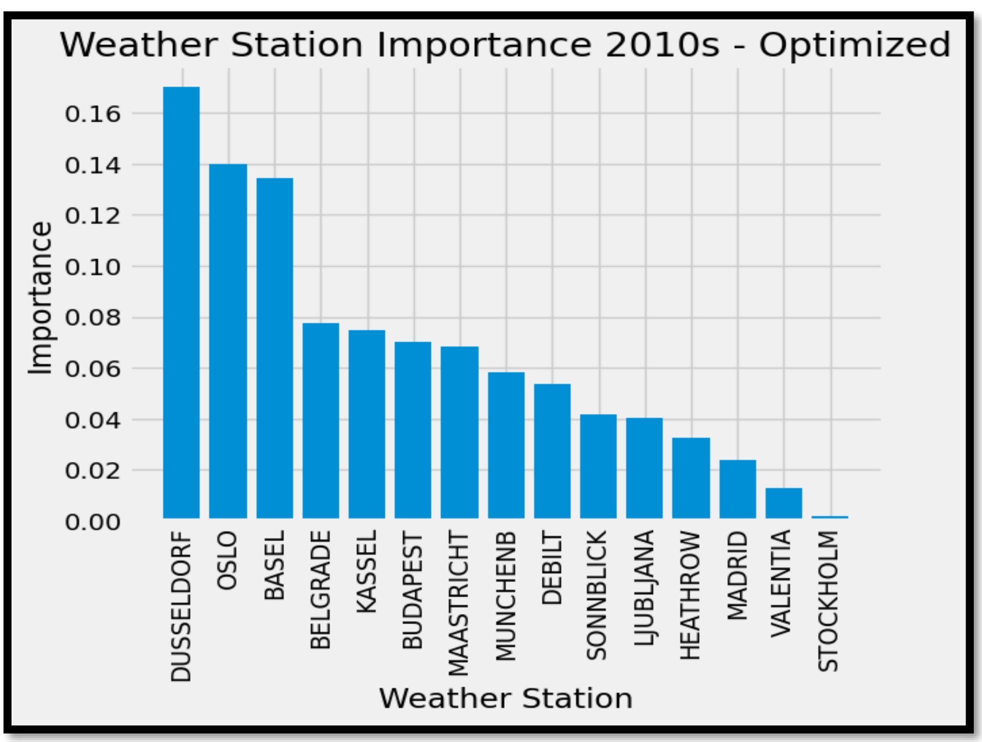
Optimized Random Forest for ClimateWins for the decade 2010-2019



Before optimization, the top stations, in descending order are Stockholm, Belgrade and Kassel (see the diagram hereunder titled ‘Weather Station Importance 2010s).

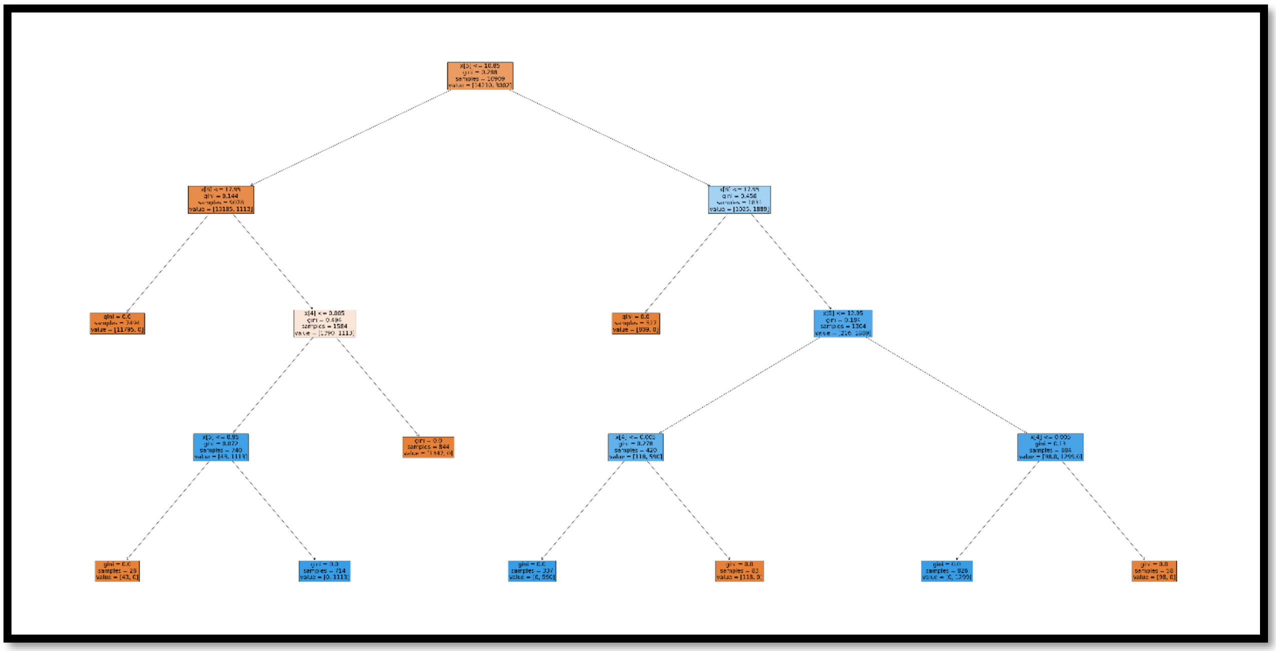


Interestingly, after optimization, the important stations have changed too as shown hereunder. The top three stations are now Dusseldorf, Oslo, and Basel.

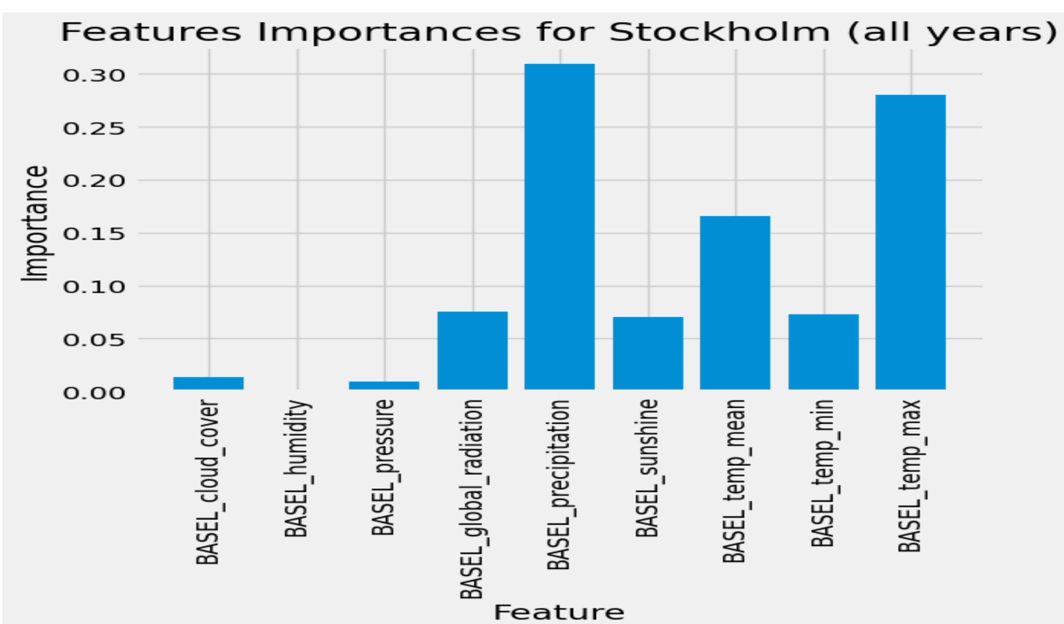


The top station before optimization, Stockholm is analyzed separately. The RF and the important features are shown hereunder.

Stockholm RF results



Important features for Stockholm



The importance of precipitation, temp_max, and temp_mean (in descending order) remain important in defining the effects of climate change.

Part 2: Deep Learning

The accuracy in Exercise 2.2 was 62.13% for RNN model and 64.35% for the CNN model. The losses were 24.5205 and nan, accordingly. Also, exercise 2.2 shows 14 weather stations in the confusion matrix. However, the accuracy in Exercise 2.4 increased to 91.90%, with an associated loss of 0.2382 and all 15 weather stations showing in the confusion matrix. The table below shows the comparison between the CNN model in exercise 2.2 against the deep learning results from exercise 2.4.

	Hyperparameters in 2.2	**Hyperparameters in 2.4**
Epochs	50	47
Batch size	550	460
N_hidden	1000	61
Kernel_size	2	1.9444298503238986
Activation	Relu	Softsign
Layers 1	/	1
Layers 2	/	2
Normalization	/	0.770967179954561
Dropout	/	0.7296061783380641
Dropout Rate	/	0.19126724140656393

The confusion matrix from Exercise 2.2 is shown hereunder hereunder.

```
[331]: # Evaluate

print(confusion_matrix(y_test, model.predict(X_test)))
```

```
180/180 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
Pred         BASEL
True
BASEL         3682
BELGRADE      1092
BUDAPEST       214
DEBILT          82
DUSSELDORF      29
HEATHROW        82
KASSEL          11
LJUBLJANA       61
MAASTRICHT       9
MADRID         458
MUNCHENB         8
OSLO             5
STOCKHOLM        4
VALENTIA         1
```

The confusion matrix for exercise 2.4 is presented hereunder showing all the 15 weather stations.

Pred	BASEL	BELGRADE	BUDAPEST	DEBILT	DUSSELDORF	HEATHROW	KASSEL
True							
BASEL	3495	93	8	15	7	8	2
BELGRADE	110	871	35	13	2	10	0
BUDAPEST	33	5	110	33	2	9	0
DEBILT	7	1	1	64	3	4	0
DUSSELDORF	4	0	0	5	10	4	1
HEATHROW	13	0	0	2	1	52	0
KASSEL	2	0	0	0	0	0	5
LJUBLJANA	8	1	2	0	0	0	0
MAASTRICHT	2	0	0	0	0	1	0
MADRID	25	3	7	0	0	10	0
MUNCHENB	5	1	0	0	0	0	0
OSLO	0	0	0	0	0	0	0
STOCKHOLM	1	0	0	0	0	0	0
VALENTIA	1	0	0	0	0	0	0

Pred	LJUBLJANA	MAASTRICHT	MADRID	MUNCHENB	OSLO
True					
BASEL	3	1	48	1	1
BELGRADE	12	0	37	0	2
BUDAPEST	7	0	15	0	0
DEBILT	1	0	1	0	0
DUSSELDORF	1	0	4	0	0
HEATHROW	1	0	12	0	1
KASSEL	1	0	2	1	0
LJUBLJANA	42	0	7	0	1
MAASTRICHT	0	4	2	0	0
MADRID	1	1	410	0	1
MUNCHENB	0	0	0	2	0
OSLO	0	0	0	0	5
STOCKHOLM	0	0	0	1	2
VALENTIA	0	0	0	0	0

Part 3

Question: write out how you might break the data down into smaller components to test and iterate upon.

Weather data is most important if it is grouped according to key/established climates and the geographical location/regions. Accordingly, weather conditions in different stations can be grouped according to Mediterranean, Tropical, Temperate. Etc. This easily groups different weather stations into cold, hot, dry, and wet regions. This also makes it easier to understand the mostly likely weather patterns.

After establishing the climates, it is also critical to focus on the seasons. Some climates have 2 seasons in a year, and others have 4 or less. These seasons control the likely weather in certain months of the year. Accordingly, the most likely weather conditions for any region can be predicted by merely knowing the climate and the month of the year which defines the season. The temperatures, precipitation, cloud cover, humidity etc goes with the established season of a given climate. However, this might still require caution as climate change is altering the known weather and climate patterns.

Question: Which model would you use for each iteration? Expand on your observations from the random forest and deep learning models.

My model of choice depends on the quantum of data that I will be dealing with. For modest amount of data, I would choose the Random Forest Model as it can handle many facets of the climate data at once (temperature, precipitation, cloud cover, humidity, etc) and can be used to predict pleasant weather conditions. Most importantly, by averaging outcomes from different decision trees, the model can estimate the interaction between these weather variables and determine the outcome weather conditions, reliably. However, when the need to address larger data sets arises, deep learning becomes ideal. By fine tuning the different estimation hyperparameters, deep learning can optimize results.

Question: What variables would you recommend that Air Ambulance pay the most attention to while deciding whether it's safe to fly?

Based on weather variables that affect safety of flights, the following are important for Air Ambulance:

i. Precipitation
ii. Wind speed
iii. Cloud cover, and
iv. Humidity

End