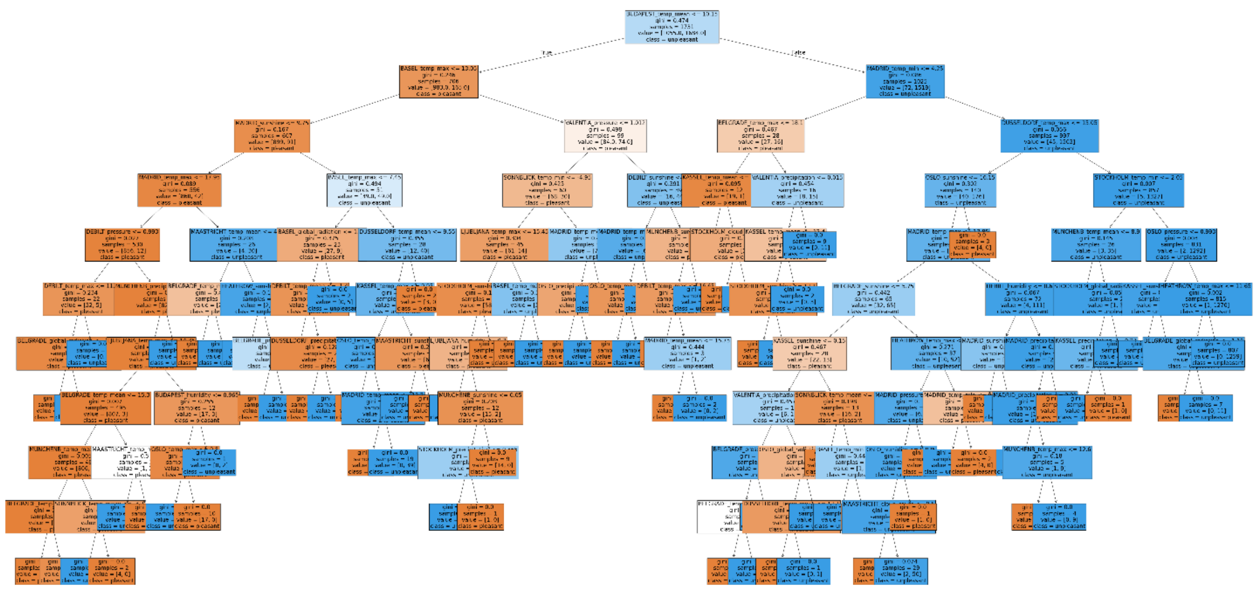
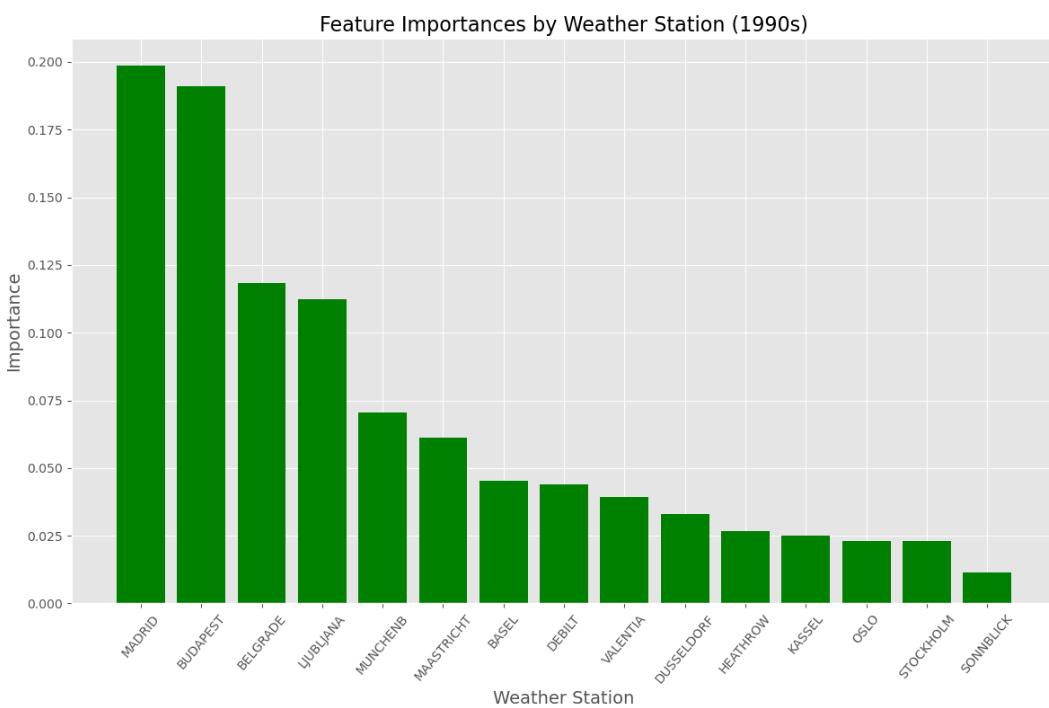
Josh Wattay
Machine Learning with Python
Task 2.4
CareerFoundry

Part 1 – Random Forest

Data	Accuracy (2.3)	Accuracy – optimized (2.4)
All weather stations 2000-2009	95.5%	95.95%
Budapest (all years available)	99.7%	100%

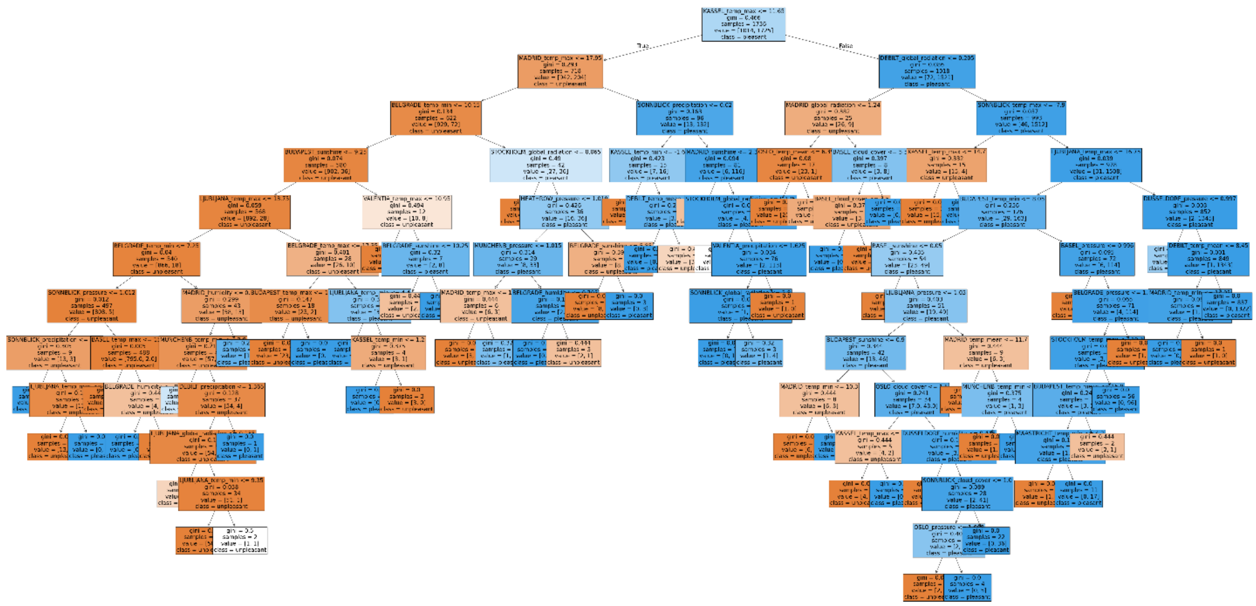
All Stations (pre-optimization)

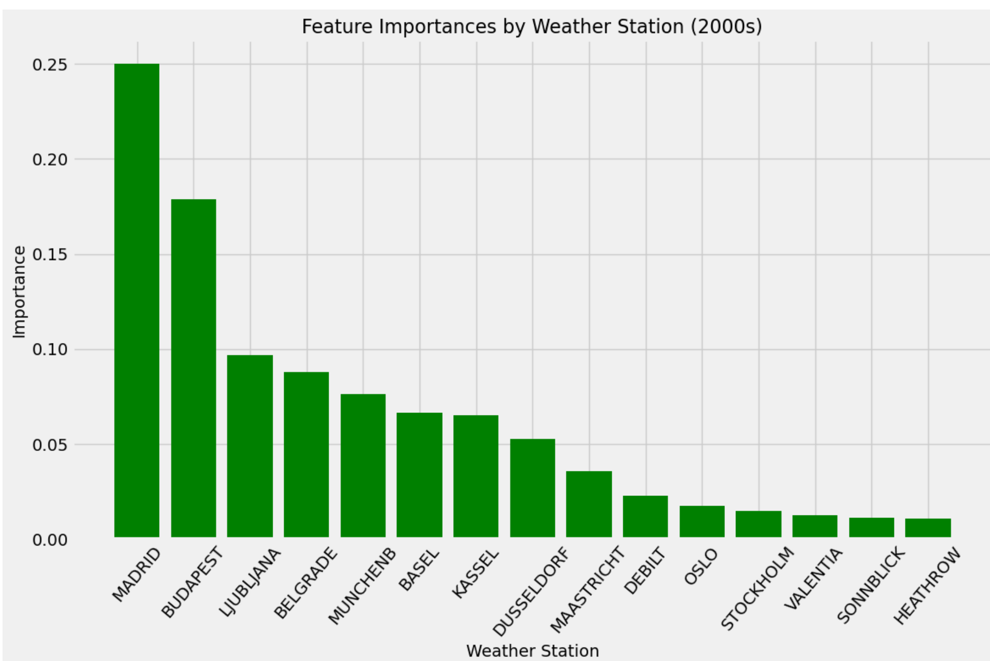




Josh Wattay
Machine Learning with Python
Task 2.4
CareerFoundry

All Stations after optimization



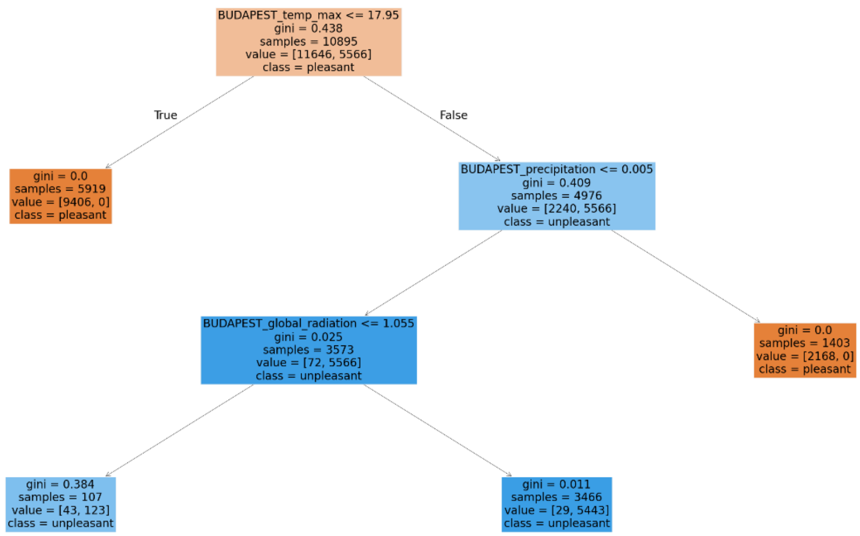


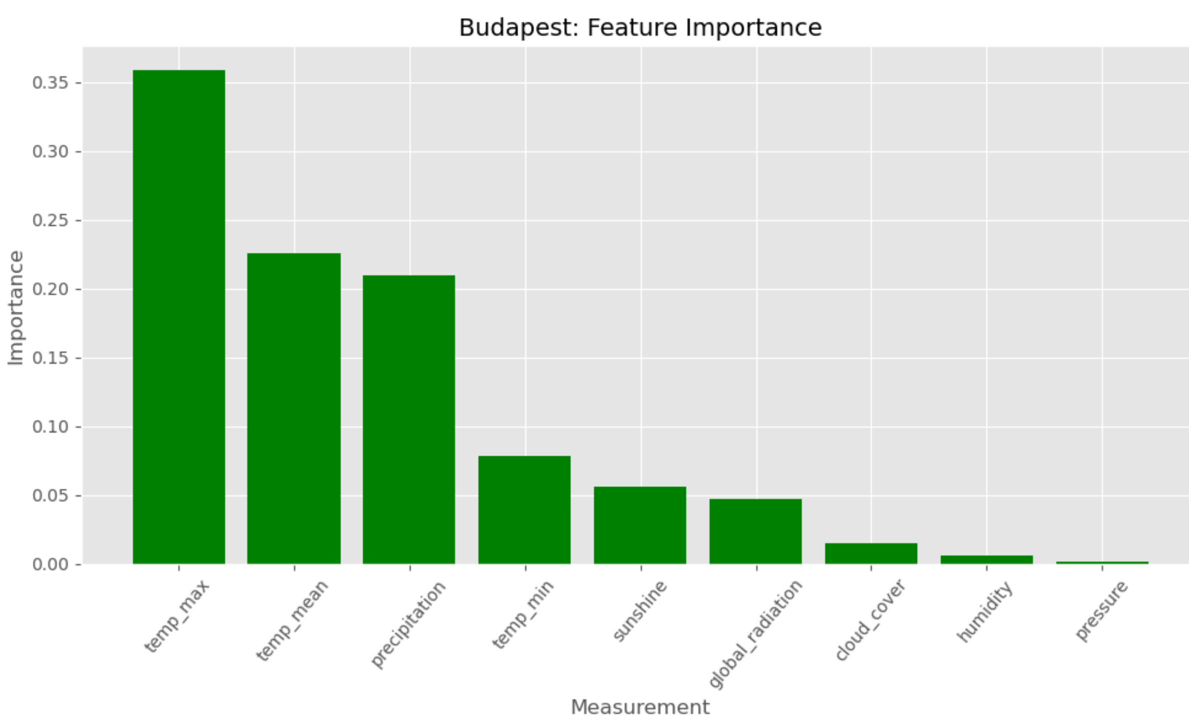
While the model's accuracy remains similar before and after hyperparameter optimization, the structural differences between the two versions highlight important changes. The optimized random forest model has more branches on the "unpleasant weather" side, whereas the unoptimized version shows more balanced branches. This shift is significant because it affects how the model prioritizes different weather patterns. Additionally, the optimized model better distinguishes the top three weather stations. In the unoptimized version, Budapest and Madrid

Josh Wattay
Machine Learning with Python
Task 2.4
CareerFoundry

have comparable influence, with all stations contributing more evenly. However, the optimized model ranks Madrid as the most influential, followed by Budapest and Ljubljana, providing clearer insights into station relevance.

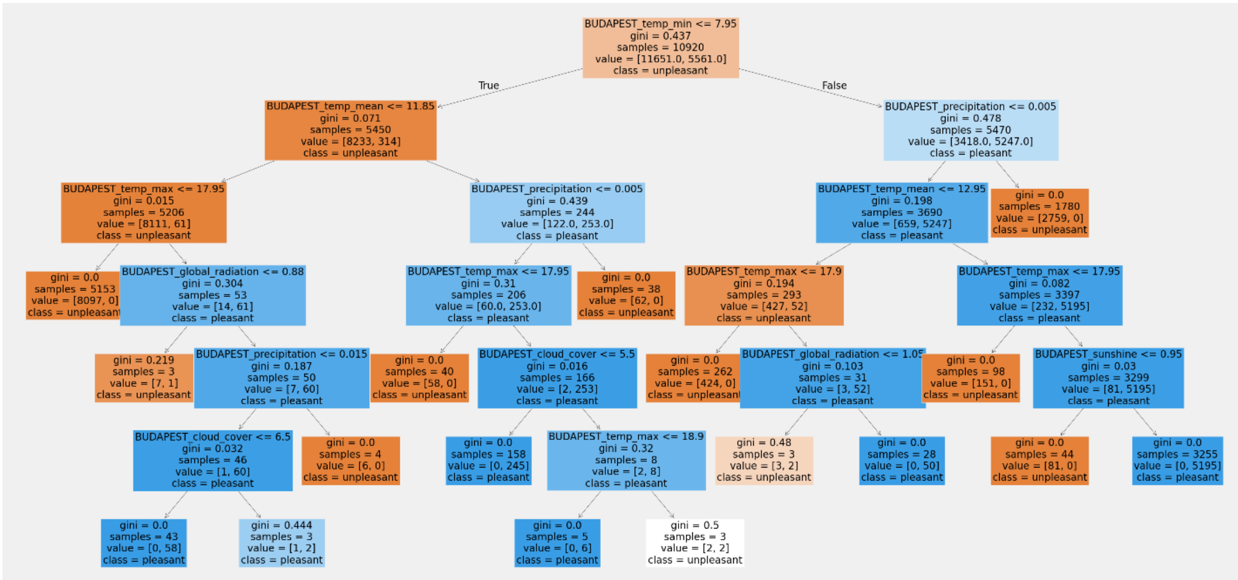
Budapest before optimization

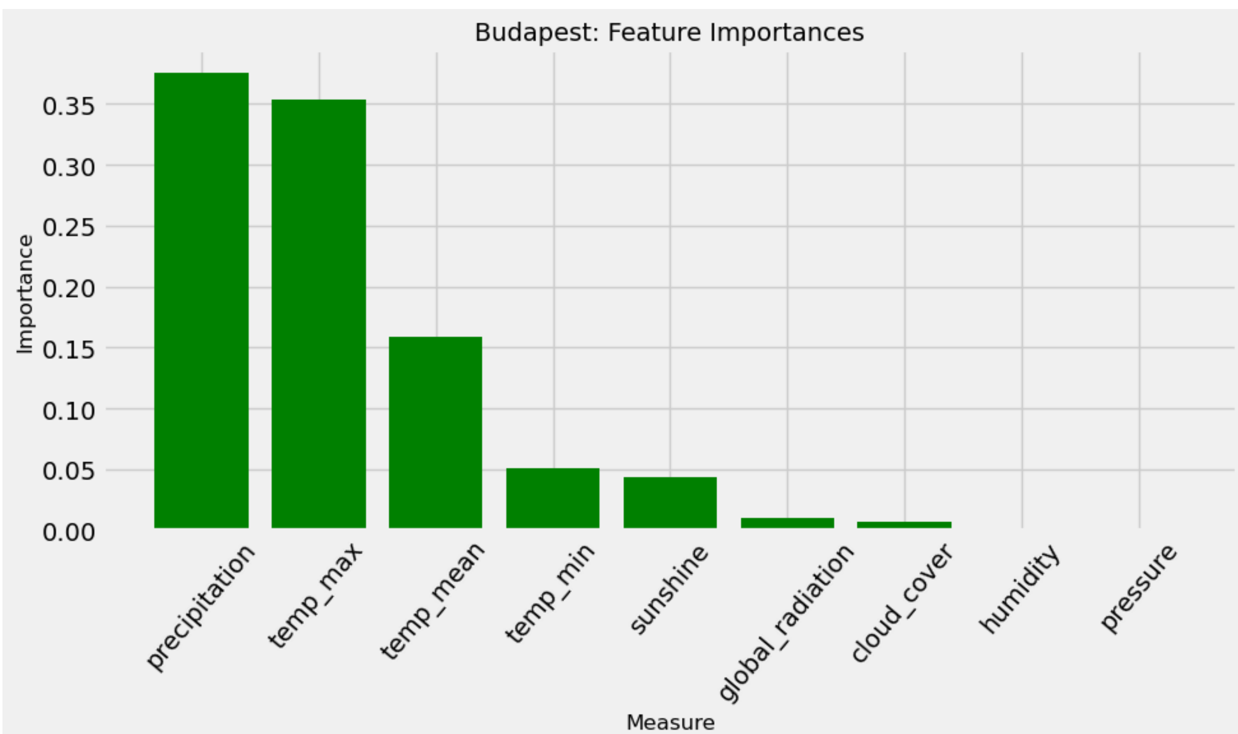




Josh Wattay
Machine Learning with Python
Task 2.4
CareerFoundry

Budapest after optimization





The differences between these images highlight the impact of optimizing the random forest model using a random search. The first set of images compares the random forest decision trees before and after optimization. In the unoptimized model, the first split is based on "temp_max," suggesting that it is the primary factor influencing the model's decisions. After optimization, the decision tree becomes more complex, with "temp_mean" taking precedence as the key decision point, indicating a shift in how the model evaluates conditions for classifying pleasant or unpleasant weather. The optimized model also shows more intricate branches, reflecting a deeper analysis of the weather data.

The second set of images focuses on feature importance for Budapest. In the unoptimized model, "temp_max" is the most influential factor, followed by "temp_mean" and "precipitation." After optimization, "precipitation" surpasses "temp_max" as the most important feature, and "temp_mean" drops in significance. This shift suggests that the optimized model places greater weight on precipitation when predicting pleasant or unpleasant weather conditions, potentially leading to different insights.

These changes are important because they show how the optimization process can alter the model's focus and decision-making criteria, potentially improving prediction accuracy or revealing new insights into the factors that drive weather classifications.

Part 2 – Deep Learning

Accuracy (2.2)	**Accuracy after Optimization (2.4)**
12.72%	94.95%

Parameter	**Parameter in 2.2**	**Hyperparameters in 2.4**
Epochs	32	38
Batch Size	64	284
Optimizer	Adam	Adadelta
Neurons / N Hidden	128	86
Kernel		2
Activation	Softmax	elu
Learning Rate		0.3892
Layers1		2
Layers2		2
Normalization		0.4055
Dropout		1.0
Dropout Rate		0.4201

Josh Wattay
Machine Learning with Python
Task 2.4
CareerFoundry

Pred / True	BASEL	BELGRADE	BUDAPEST	DEBILT	DUSSELDORF	HEATHROW	KASSEL	\
BASEL	3525	42	10	3	3	7	0	
BELGRADE	34	1023	3	5	0	4	1	
BUDAPEST	11	5	172	3	1	2	1	
DEBILT	4	1	6	73	1	3	1	
DUSSELDORF	6	0	0	3	30	3	0	
HEATHROW	5	2	0	3	0	80	1	
KASSEL	1	1	0	0	1	0	8	
LJUBLJANA	1	1	0	0	0	2	0	
MAASTRICHT	1	0	0	0	1	0	0	
MADRID	17	1	2	0	1	13	0	
MUNCHENB	6	0	0	0	0	0	1	
OSLO	1	0	0	0	0	0	0	
STOCKHOLM	2	0	0	0	1	0	0	
VALENTIA	1	0	0	0	0	0	0	

Pred / True	LJUBLJANA	MAASTRICHT	MADRID	MUNCHENB	OSLO	STOCKHOLM	VALENTIA
BASEL	3	2	14	0	1	0	0
BELGRADE	0	0	2	1	0	0	0
BUDAPEST	1	0	9	0	0	0	0
DEBILT	0	0	0	0	0	0	0
DUSSELDORF	0	1	5	0	1	0	0
HEATHROW	5	0	5	0	0	0	0
KASSEL	2	0	3	1	0	0	0
LJUBLJANA	44	0	3	0	0	0	0
MAASTRICHT	0	3	0	0	0	0	0
MADRID	1	0	481	0	0	0	0
MUNCHENB	0	0	0	4	0	0	0
OSLO	0	0	0	1	2	0	0
STOCKHOLM	0	0	0	0	0	1	0
VALENTIA	0	0	0	0	0	0	2

After optimizing the model, there was a significant improvement in accuracy, jumping from 12.72% in version 2.2 to 94.95% in version 2.4. The hyperparameter adjustments played a crucial role in this improvement. Notably, the optimizer was changed from Adam to Adadelta, the learning rate was fine-tuned to 0.3892, and the batch size increased from 64 to 284. Additionally, the number of neurons in the hidden layers decreased from 128 to 86, and activation functions shifted from Softmax to elu. New layers and kernels were added, with dropout regularization and normalization incorporated to further improve the model's performance. These changes resulted in a more efficient and accurate model that was better able to generalize the data.

Josh Wattay
Machine Learning with Python
Task 2.4
CareerFoundry

Part 3 – Iteration

To test and iterate on the dataset, I recommend breaking the data down into smaller components based on location, time periods, weather features, and conditions. For location, data should be separated by specific regions or weather stations (e.g., Budapest, Madrid, Ljubljana), allowing the model to focus on local weather patterns for each area. Time period segmentation, such as by year or season, would reveal insights into how weather patterns evolve and influence flight safety. Additionally, grouping data by weather variables, such as temperature, precipitation, and humidity, would highlight the importance of each feature. Lastly, creating subsets for pleasant and unpleasant weather conditions will help test the model's ability to predict outcomes based on extreme or optimal conditions.

For each iteration, I would suggest using a random forest model as the starting point, due to its ability to handle non-linear relationships and multiple weather variables effectively. Based on earlier observations, the optimized random forest model demonstrated a significant shift from focusing on "temp_max" to "precipitation" when predicting weather conditions, and the decision trees became more complex after optimization, capturing more granular data patterns. This model will allow for observing feature importance and key variables affecting flight safety. Additionally, the **deep learning model** should be considered for more complex relationships. After optimization, the deep learning model's accuracy improved drastically from 12.72% to 94.95%. Changes such as switching from the Adam optimizer to Adadelta and increasing the batch size allowed the model to handle larger datasets and identify nuanced patterns. Testing how this model adapts to different weather stations and time periods would yield valuable insights.

In terms of key variables for Air Ambulance to focus on when determining flight safety, **precipitation** should be prioritized, as it became the most important factor after optimization. Sudden changes in precipitation, such as rain or snow, could significantly affect flight safety. **Temperature** (both max and mean) is also crucial, as temperature fluctuations play a critical role in flight dynamics. Although **cloud cover** and **radiation** were lower in importance in some iterations, they remain significant as they affect visibility and flight stability, particularly during adverse weather conditions. Finally, while not explicitly mentioned, **wind speed** and **direction** are critical for air travel safety and should be factored into future models. By iterating on these smaller data subsets, using random forest for feature importance, and deep learning for complex relationships, ClimateWins models can identify the most relevant variables for ensuring flight safety, directly impacting the success of Air Ambulance.