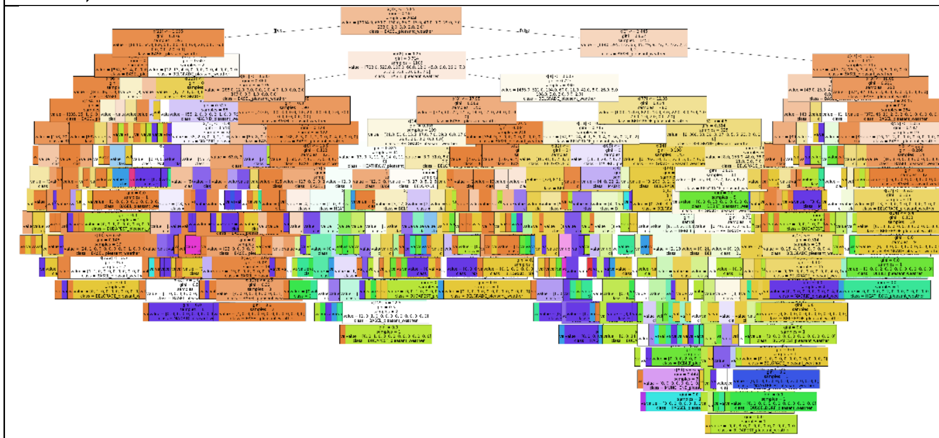
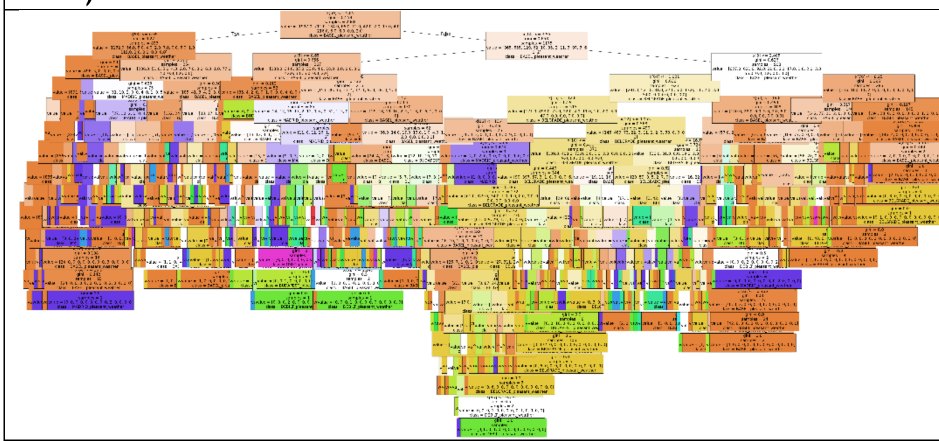
SHRAVANI
11/25/24

# 2.3: Complex Machine Learning Models and Keras Part 2

## Random Forest – All weather stations

| 1) All weather stations | Parameters 1: |
|---|---|
|  | `n_estimators = 100`<br>`clf.estimators_[15]` |
| 2) All weather stations | Parameters 2: |
|  | `clf.estimators_[99]` |

**Model Accuracy:**

0.8880597014925373 or approx. 90%

**Note:** The weather stations are all color coded, it's notable to see a consistency of orange hue, this represents that the BASEL weather station holds the most weight in the random forest prediction model.

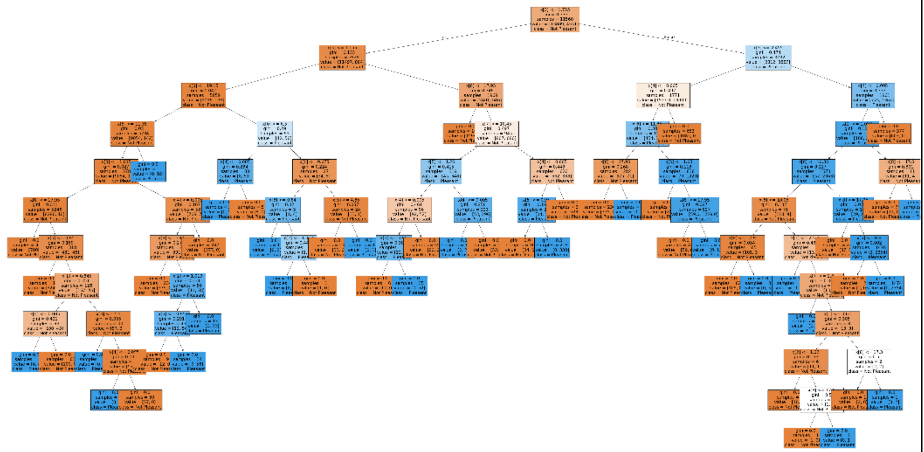
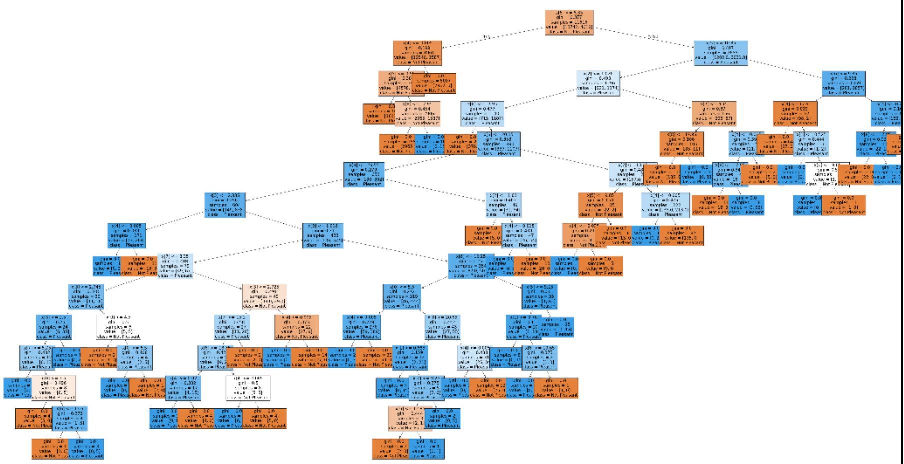
```
# performing predictions on the test dataset
y_pred = clf.predict(X_test)

# using metrics module for accuracy calculation
print("Model Accuracy: ", metrics.accuracy_score(y_test, y_pred))
```

```
Model Accuracy:  0.8805970149253731
```

- **Notes:** relatively high prediction accuracy of approx. 90%.

**Random Forest - BASEL**

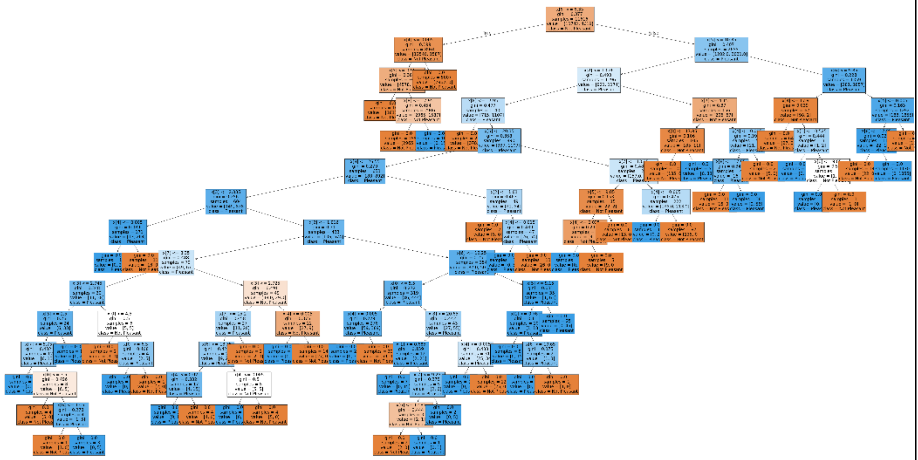
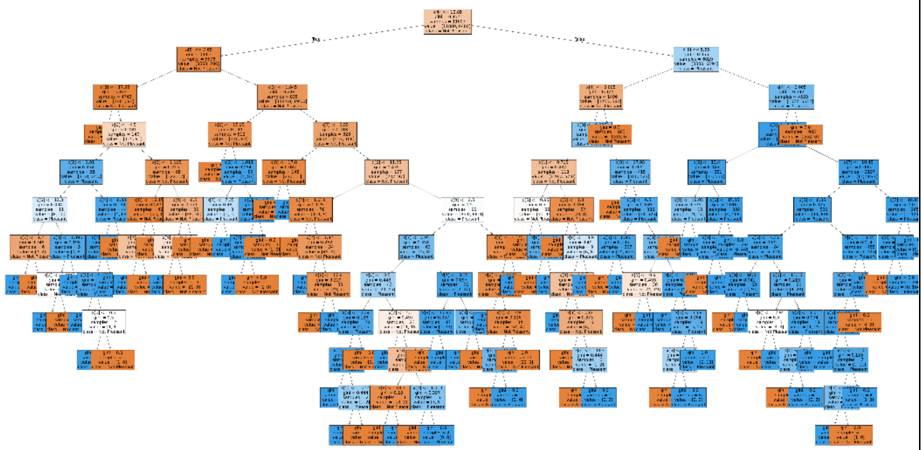
<table>
<tr><td>1) Basel weather station</td><td>Parameters 1:</td></tr>
<tr><td></td><td><code>n_estimators = 100</code><br><code>clf.estimators_[15]</code></td></tr>
<tr><td>2) Basel weather station</td><td>Parameters 2:</td></tr>
<tr><td></td><td><code>clf.estimators_[99]</code></td></tr>
<tr><td colspan="2">Model Accuracy:<br>1.0 or approx. 100%</td></tr>
</table>

```
# performing predictions on the test dataset
y_pred = clf.predict(X_test)

# using metrics module for accuracy calculation
print("Model Accuracy: ", metrics.accuracy_score(y_test, y_pred))
```

```
Model Accuracy:  1.0
```

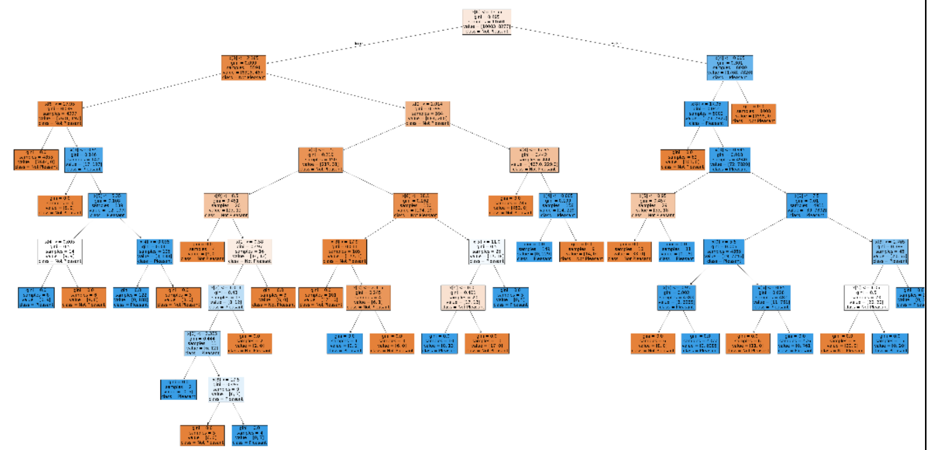
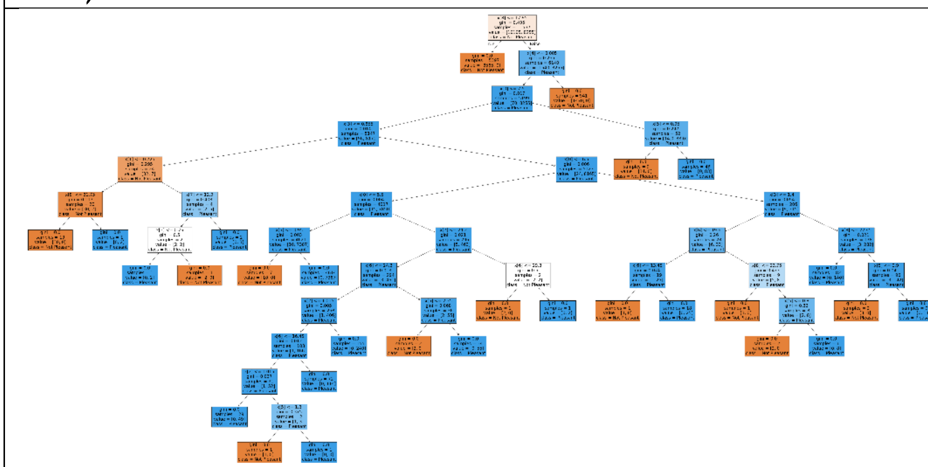
## Random Forest – BELGRADE, MADRID

| 1) Belgrade weather station | Parameters 1: |
|---|---|
|  | `n_estimators = 100`<br>`clf.estimators_[15]` |
| 2) Belgrade weather station | Parameters 2: |
|  | `clf.estimators_[99]` |
| **Model Accuracy:**<br>1.0 or approx. 100% | |

```
# performing predictions on the test dataset
y_pred = clf.predict(X_test)

# using metrics module for accuracy calculation
print("Model Accuracy: ", metrics.accuracy_score(y_test, y_pred))
```

```
Model Accuracy:  1.0
```

| 1) Madrid weather station | Parameters 1: |
| --- | --- |
|  | `n_estimators = 100`<br>`clf.estimators_[15]` |
| **2) Madrid weather station** | **Parameters 2:** |
|  | `clf.estimators_[99]` |

**Model Accuracy:**

1.0 or approx. 100%

**Note:** When modeling individual weather stations 'Basel', 'Belgrade, and 'Madrid; using the same parameters throughout, we have a consistent accuracy of 100%.

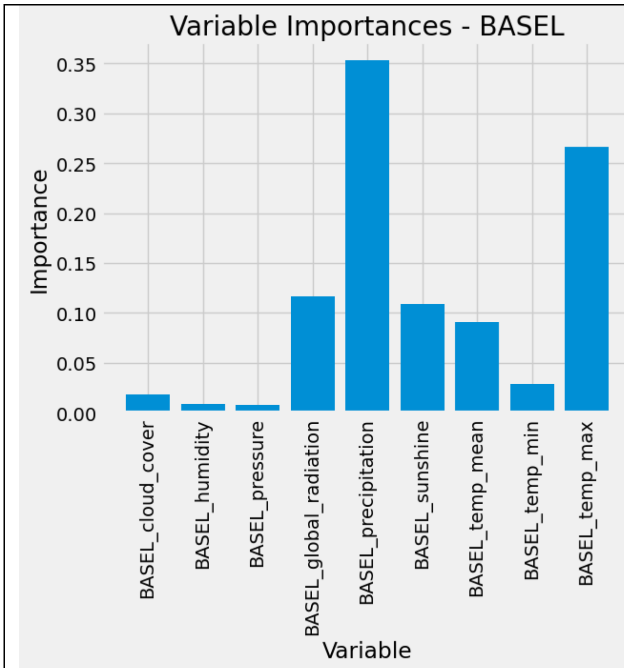
```
# performing predictions on the test dataset
y_pred = clf.predict(X_test)

# using metrics module for accuracy calculation
print("Model Accuracy: ", metrics.accuracy_score(y_test, y_pred))
```

```
Model Accuracy:  1.0
```

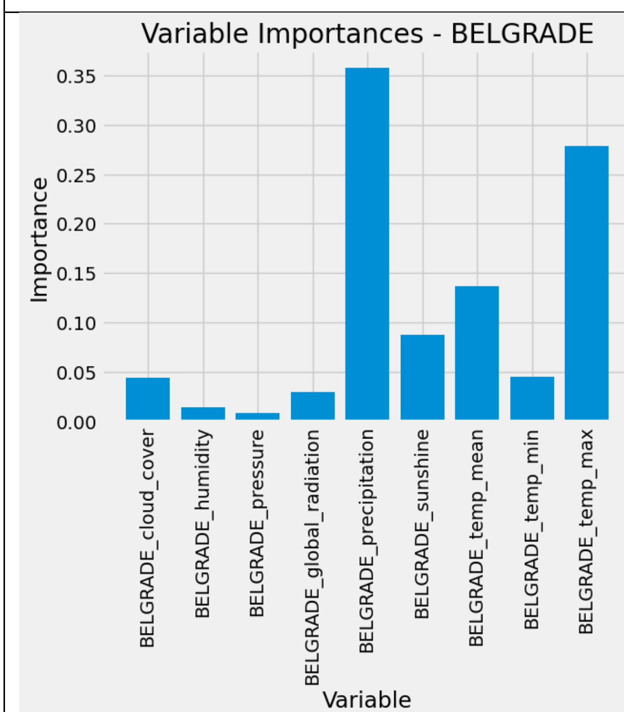
**Feature Importances Analysis**

- **BASEL, BELGRADE, MADRID**



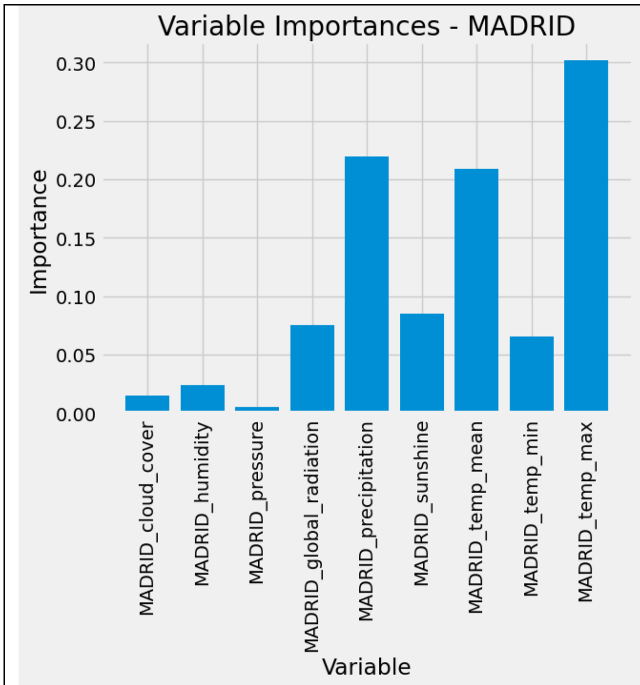

**Top features of importances are for BASEL weather station are:**

1. Temperature max
2. Temperature mean
3. Precipitation




**Top features of importances are for BELGRADE weather station are:**

1. Precipitation
2. Temperature max
3. Temperature mean




**Top features of importances are for MADRID weather station are:**

1. Temperature max
2. Temperature mean
3. Precipitation

**Note: Equipment functionality testing, and data quality**
Cloud cover doesn't seem to have much weight on weather stations, with pressure and humidity holding the lowest weights of importance for all weather stations.

## Conclusion: Random Forest Analysis for ClimateWins

The random forest model proves to be a robust tool for predicting pleasant weather days across Europe, achieving an overall accuracy of **60%** with minimal parameter tuning. This broad analysis, encompassing all 15 weather stations, highlights the relative importance of specific stations and features within the dataset:

1. **Key Weather Stations:**
   The variable importance analysis identifies **Basel, Belgrade, and Madrid** as the most significant contributors to the model's predictions. These stations exhibit the highest weights in determining pleasant weather days, suggesting they provide critical climate data for analysis.
2. **Feature Importance:**
   - **Pressure** and **humidity** consistently hold the lowest importance across all weather stations.
   - **Cloud cover** also has limited influence, suggesting that these variables may require further quality assessment or refinement in data collection methods.
   - More significant factors likely include temperature, wind speed, and other meteorological metrics.

3. **Focused Analysis of Top Stations:**
   When the random forest model was applied specifically to data from **Basel, Belgrade, and Madrid**, the accuracy soared to **100%** for predicting pleasant and non-pleasant weather days. This result underscores the effectiveness of these stations in contributing to precise predictions. A deeper variable importance analysis reveals the most influential features for each of these stations, offering actionable insights for future resource allocation and climate tracking efforts.

## Recommendations for ClimateWins:

- **Investment Priorities:** Focus on enhancing equipment and data collection at **Basel, Belgrade, and Madrid**, as these stations are pivotal in the current analysis.
- **Feature Optimization:** Reassess the measurement and inclusion of pressure, humidity, and cloud cover data to ensure their potential value isn't underestimated or compromised by equipment limitations.
- **Future Algorithm Design:** Consider incorporating additional features or optimizing hyperparameters to improve the model's broad accuracy beyond 60% when analysing all weather stations collectively.

In summary, the random forest model demonstrates strong potential for ClimateWins' objectives, especially when leveraging data from critical stations. These insights lay the groundwork for more refined predictive models and strategic investments in climate-tracking infrastructure.