**Course Code and Title:**



**Bachelor of ICT Assessment Cover Sheet**

Complete and attach this cover sheet to your assessment before submitting

202001648 & 201900500

**Assessment Title:**

**Learning Outcomes:**

LO1 – Demonstrate critical knowledge and understanding of the fundamental concepts and techniques of Data Mining

LO2 – Utilize data mining software to solve defined and undefined problems.

LO3 – Interpret and evaluate obtained numerical and graphical data to recommend the best model to be used for prediction purposes

Group Project

IT8416 – Data Mining

**Student IDs:**

**Student Names:**

**Tutor:**

Talal Alaamer & Amira Alawadhi

**The maximum grade granted for late submission is 60 % for up to 3 calendar days. A grade of 0 will be allocated for submission after 3 days**

**Late Rule:**

Group 2

**Group No:**

**Sunday**, 4th June 2023 by 11:55 p.m.

**Group Project Submission Date**

Sini Raj Pulari

#### By submitting this assessment for marking, either electronically or as hard copy, I confirm the following:

* *This assignment is* ***our own group work***

##### Any information used has been properly referenced.

##### We understand that a copy of my work may be used for moderation.

##### We have kept a copy of this assignment

Do not write below this line. For Polytechnic use only.

**Assessor:**

**Date of Marking:**

**Grade/Mark:**

# Table of Contents

[Task 1 3](#_Toc125276489)

[Task 2 3](#_Toc813796762)

[Task 3 3](#_Toc2060020189)

[Task 4 3](#_Toc116813135)

[Task 5 3](#_Toc1778500741)

[Task 6 3](#_Toc1370302581)

[Reflection 4](#_Toc1796266621)

[Roles & Logs 4](#_Toc1020450104)

# Task 1

A weather forecast's business goals are to assist companies in making safe decisions about the weather. All businesses worldwide will gain from precise weather prediction knowledge since it allows them to plan and be ready for everything. For instance, a construction firm may require weather forecast knowledge to plan its work schedule and guarantee the safety of its employees. Like how a shop arranges inventory and marketing tactics using weather prediction.

Weather forecasting requires much observation and data processing, making it a complicated and sometimes difficult skill. Data mining techniques will help ease this process.

The objective is to examine vast volumes of data and find patterns that may be utilized to produce more precise forecasts. Data mining can be used to enhance weather forecasting. It may be used to study past weather data and spot trends that can be used to forecast weather conditions in the future. To provide more precise short-term weather forecasts, data mining may also be utilized to analyze real-time weather data and find trends.

Some drawbacks of weather forecasting include the following:

* It is incredibly challenging to predict the weather accurately.
* The cost of keeping track of so many variables from many sources is high.
* The cost of the computers required to do the millions of computations required is high.
* If the weather does not match the forecast, the weather forecasters are held responsible.

However, despite these challenges and with the advancement of technology and computers, weather forecasting has come a long way in recent years and has become more accurate than ever before.

Through the analysis of huge amounts of data and the discovery of patterns, predictive data mining techniques may be utilized to forecast the weather with greater accuracy. Decision trees, neural networks, and support vector machines are a few of the methods that may be applied.

This project aims to produce more accurate weather forecasts using previous data to find patterns and therefore infer from it and get more accurate weather predictions.

# Task 2

The data source that we have chosen for our project is a weather prediction dataset taken from the ‘European Climate Assessment & Dataset’ and can be found on this link <https://zenodo.org/record/4770937#.ZFerMHZByUl>.

We have chosen this dataset for the following reasons:

- It is relevant to our issue.

- It has many different attributes which can help get more accurate results

- It is not too big to process and analyze, but big enough to catch the patterns and trends.

- Most of the data is accurate, it only has a few mistakes such as missing numbers, duplicates, and outliers.

The dataset chosen has a total of:

* 165 Attributes (All of which are quantitative data)
* 3654 Records/Instances

It also contains some missing and redundant data.

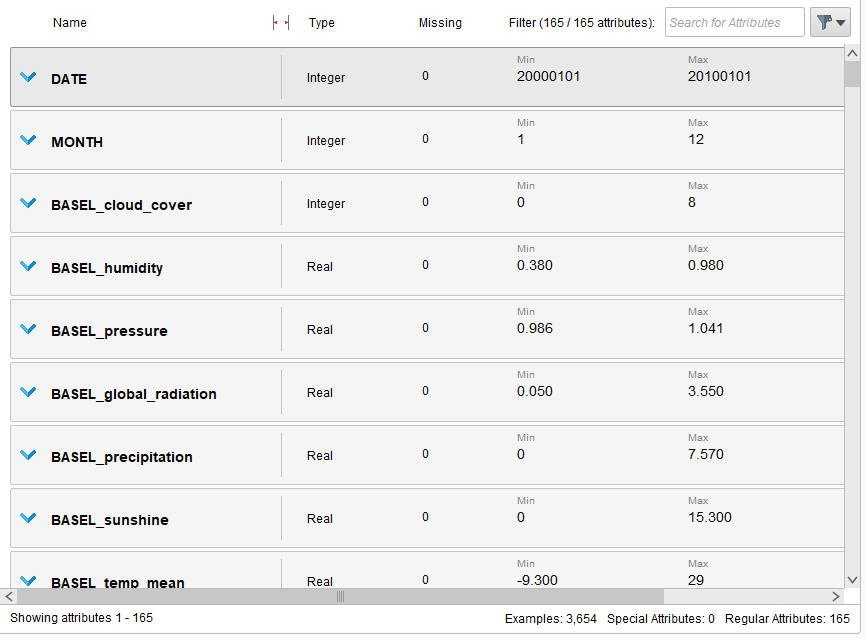


Figure 1: This figure shows some attributes of the dataset along with the total number of attributes.

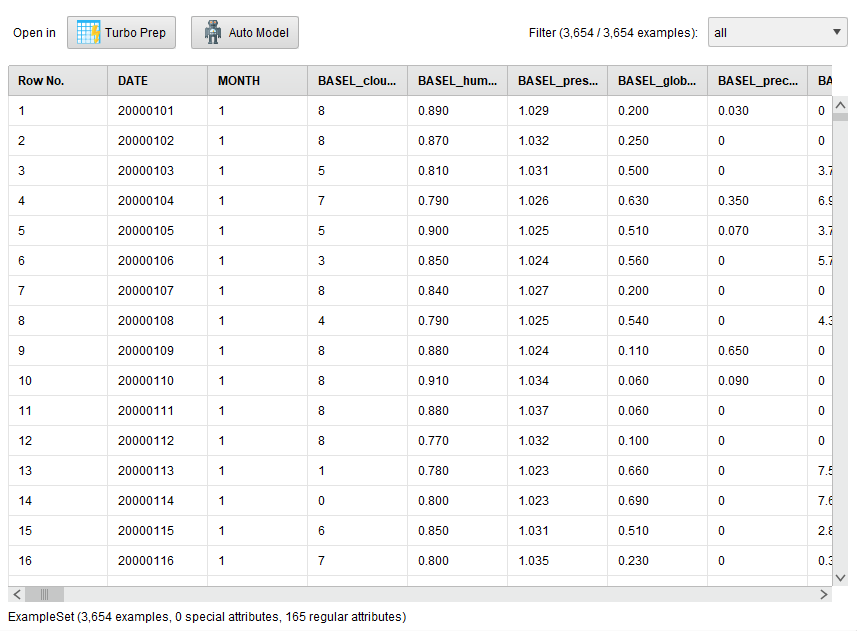


Figure 2: This figure shows some records/instances of the dataset along with the total number of records.

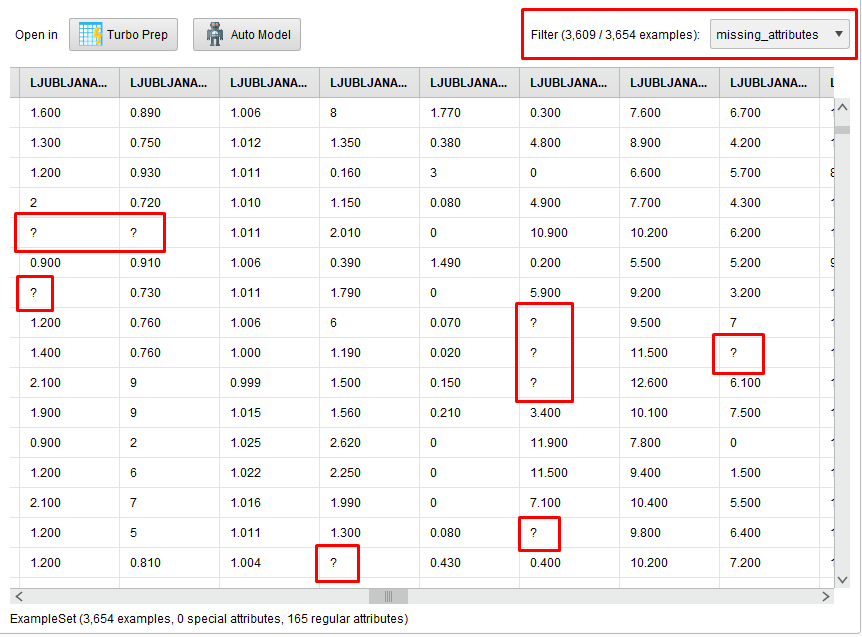


Figure 3: This figure shows some missing data found in some records. The missing data is represented by a “?”.

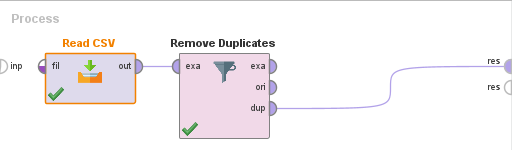


Figure 4: This figure shows the usage of “Remove Duplicates” operator in rapid miner and connecting the “dup” output port to the results to show the duplicate records.

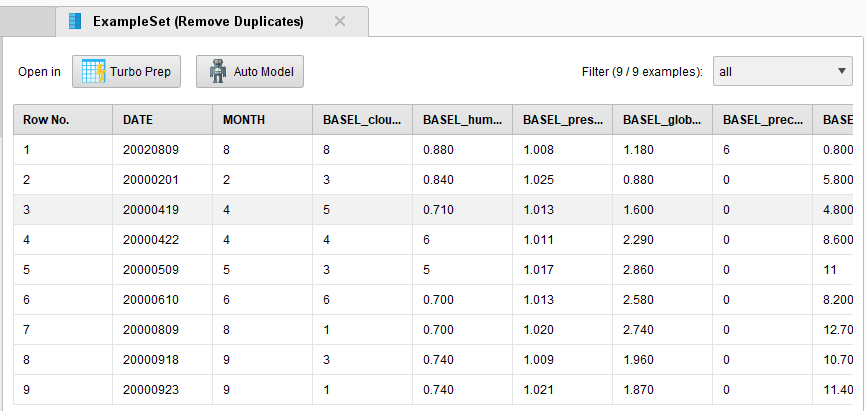


Figure 5: This figure shows the duplicate records of the data set on output.

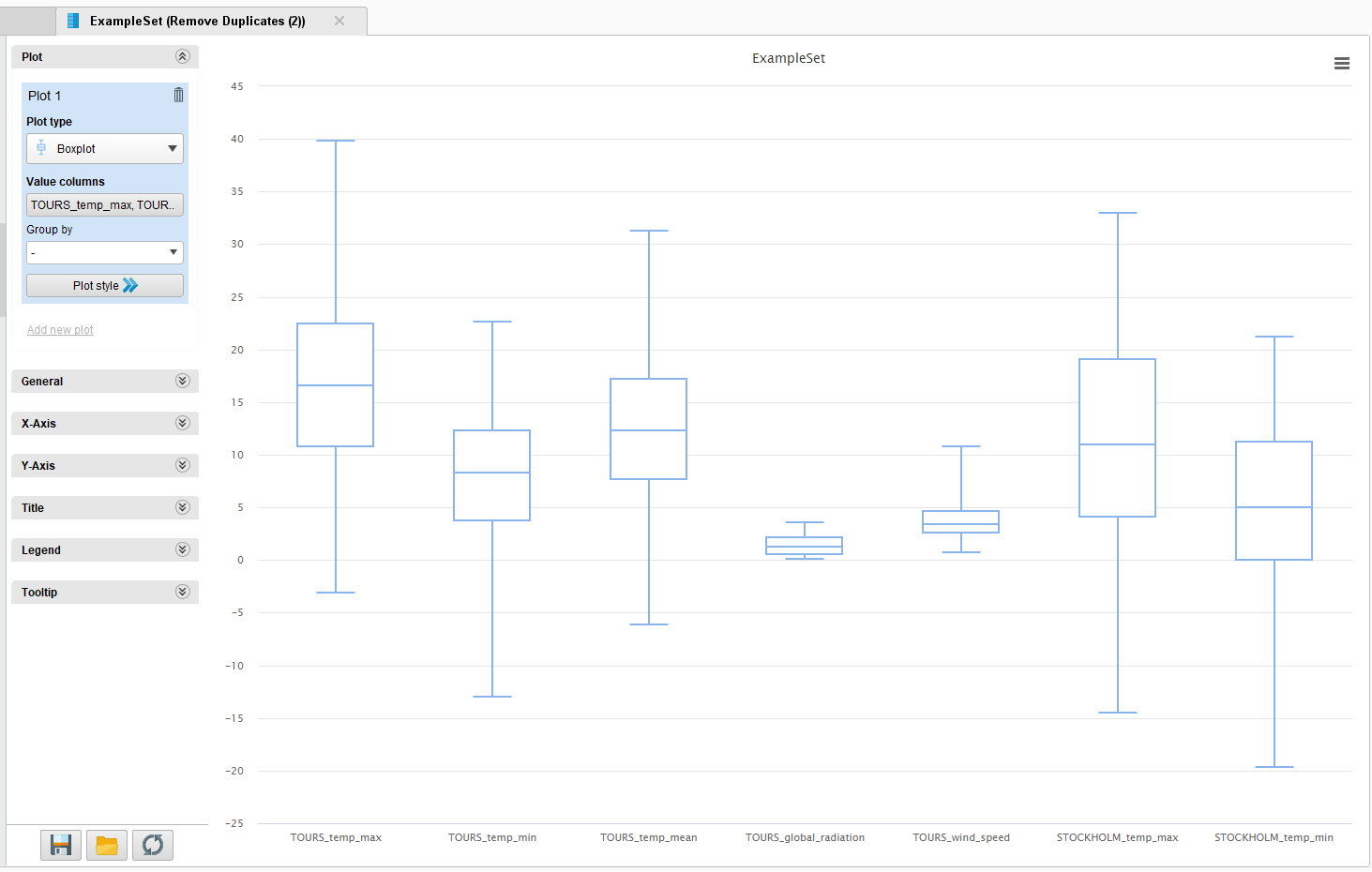


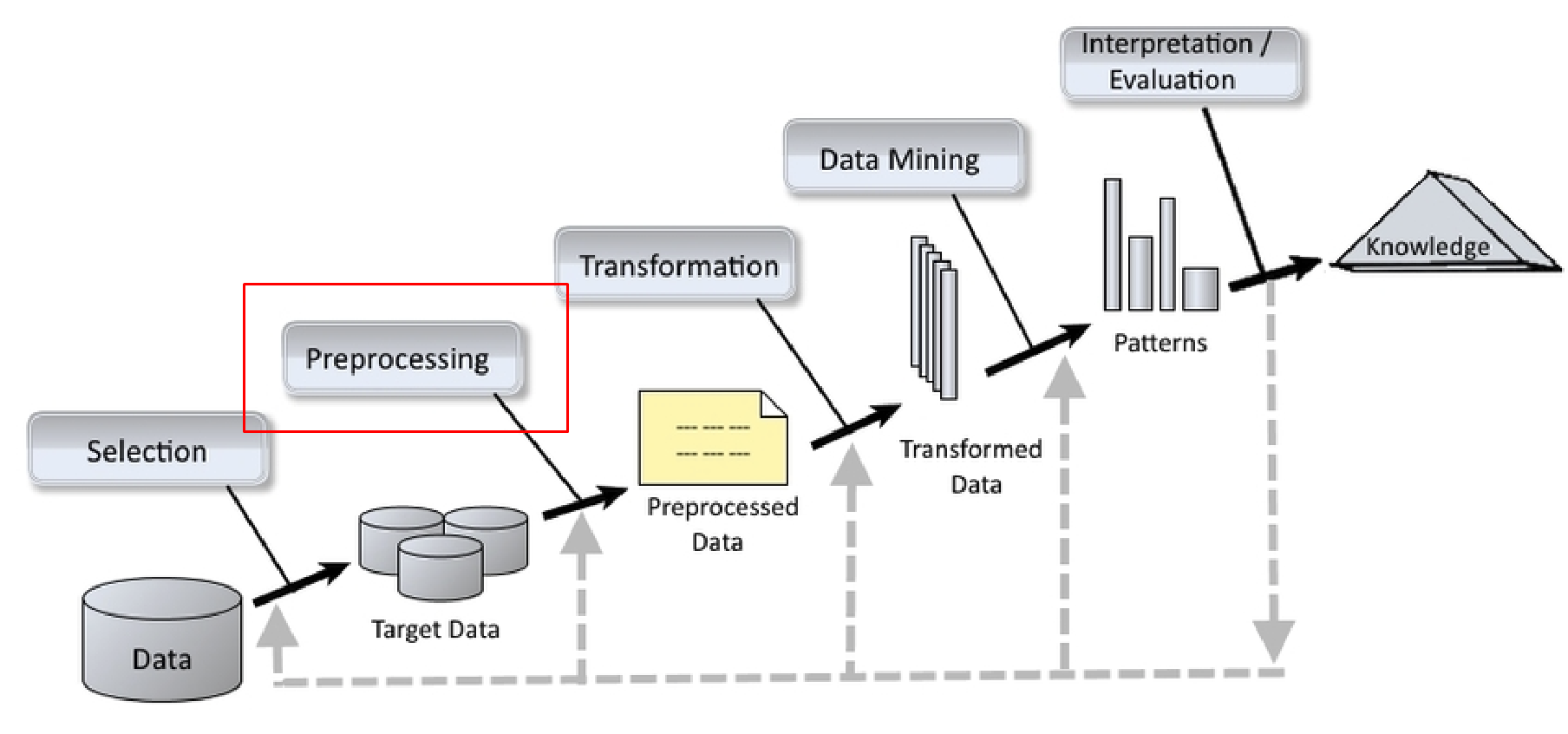
Figure 6: This figure shows the box plot diagram for some of the dataset’s attributes. It can be noticed that most of the values are positive (greater than 0) and that the range varies a lot from one attribute to another. Also, the skewness of each attribute varies but mostly it is positively skewed.



Figure 7: This figure shows the scatter matrix for some of the dataset’s attributes. This matrix helps in showing the relationship between two different attributes. Based on the figure, it can be interpreted that most of the attributes have a positive correlation with only a few exceptions.

# Task 3

Data preprocessing or data wrangling is an essential step in any data analysis or machine learning project. It involves converting raw, unclean data into a usable format that can be fed into an analytical model. It entails transforming dirty, raw data into a format that may be used as input for an analytical model. Data preprocessing may assist in increasing the data's quality and dependability, lessen noise and mistakes, improve the model's performance and accuracy, and draw out valuable insights from the data.

Figure 8: This figure shows the KDD architecture diagram for data mining. Data preprocessing should be done right after the data selection and before applying any data mining algorithms to get more accurate results.

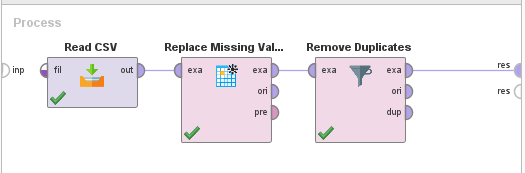
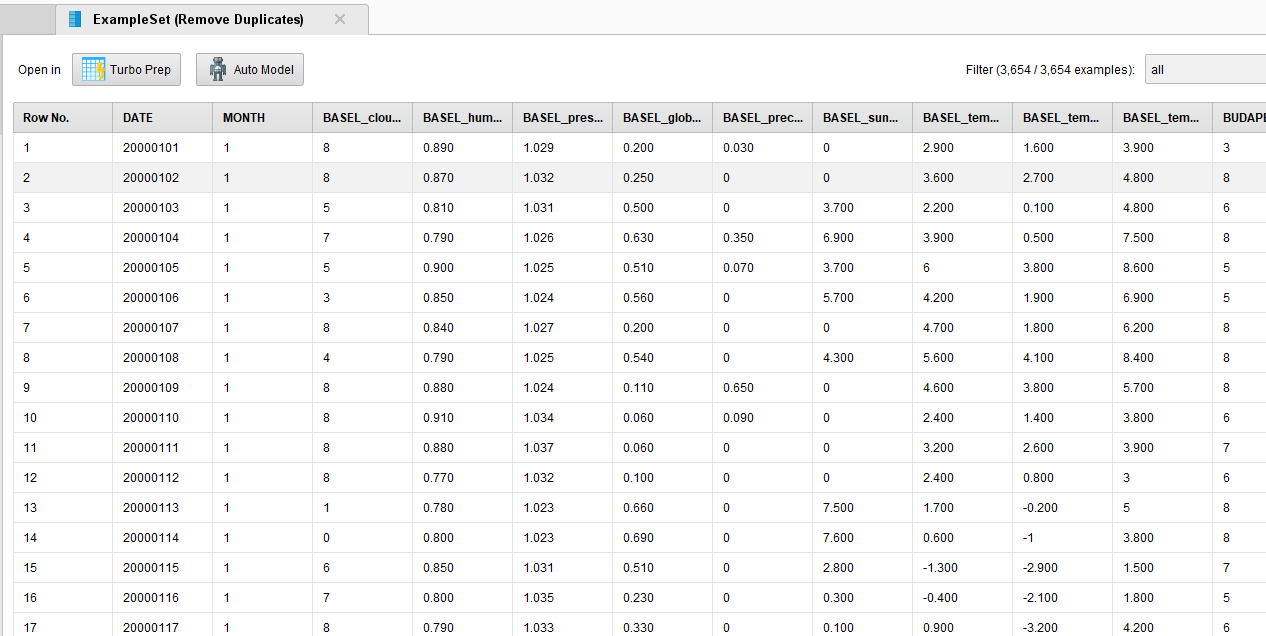


Figure 9: This figure shows the usage of the “Replace Missing Values” and “Remove Duplicates” operators to ensure that no fields are redundant or left empty.

Figure 10: This figure shows the result of removing all duplicate values and replacing all missing values with the column’s average value.

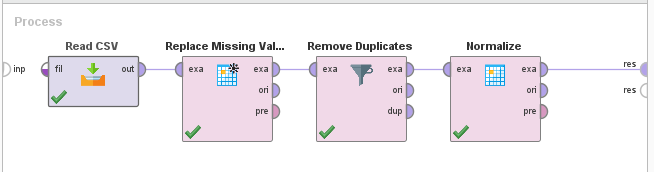


Figure 11: This figure shows the usage of the “Normalize” operator to minimize redundancy and minimize the range using Z-transformation method.

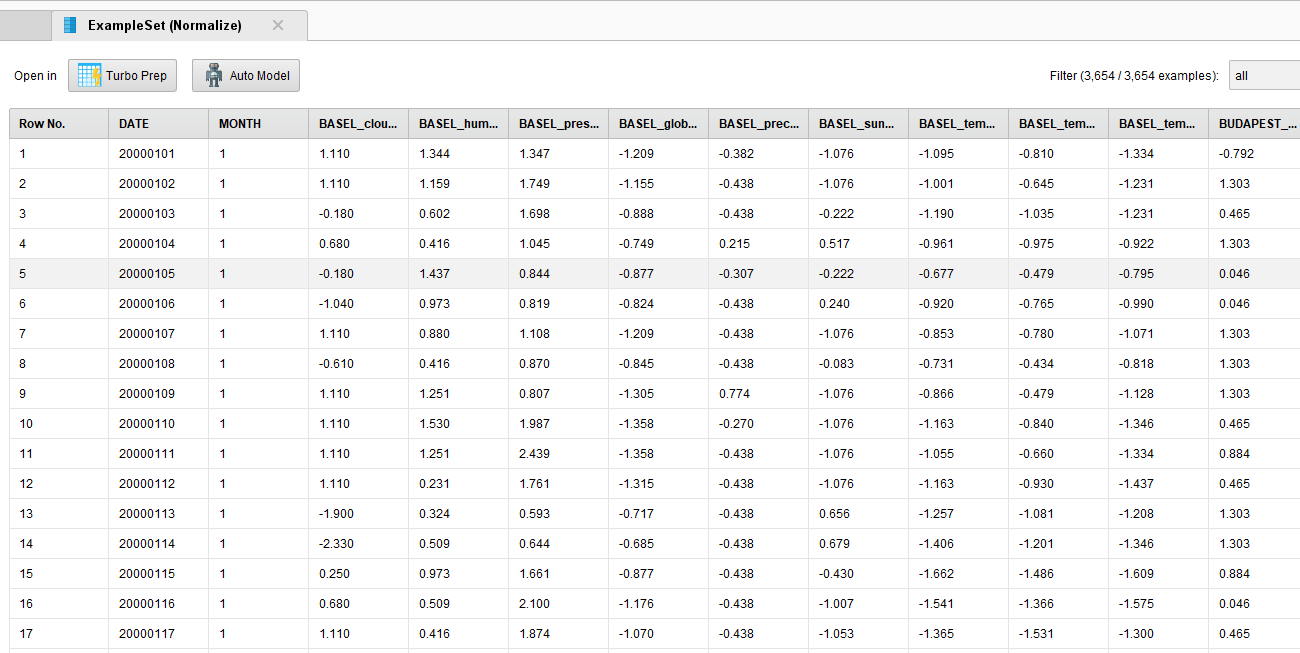


Figure 12: This figure shows the output for after the normalization process. Most of the attribute values have been reduced to a lower range.

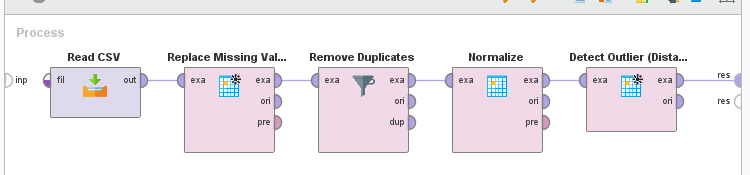


Figure 13: This figure shows the usage of the “Detect Outlier” operator. It finds the outlier values based on the distance.

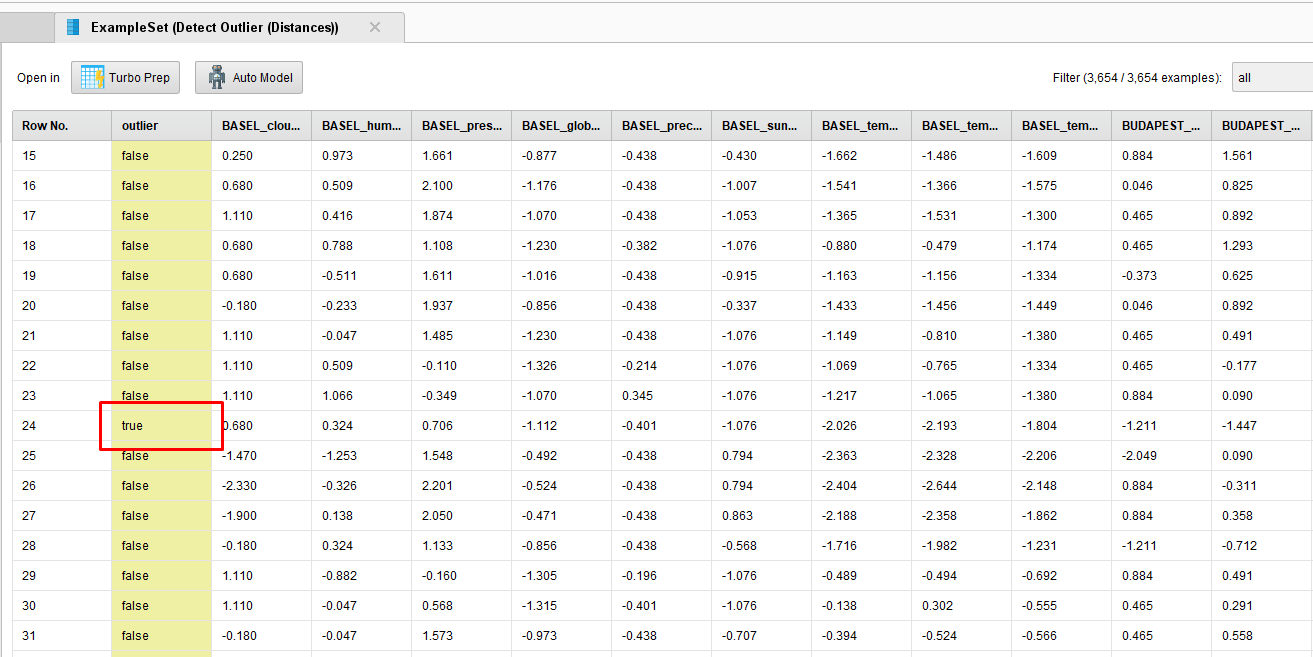
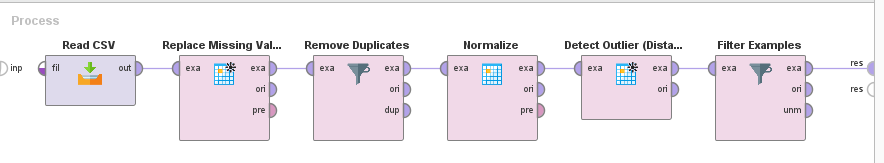


Figure 14: This figure shows the output after detecting outliers. Any outliers have been marked as “true” and therefore it would be best to remove them.



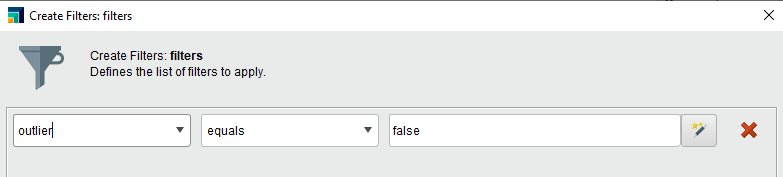
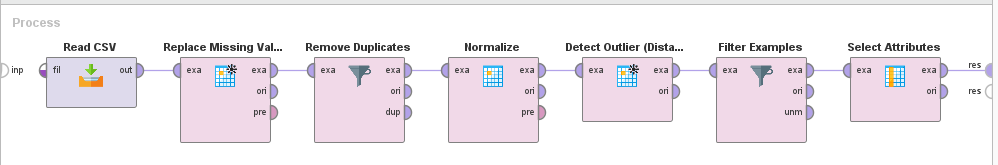


Figure 15, 16: These figures show the usage of the “Filter Examples” operator along with the specified filter condition. It keeps the records that have the outlier field’s value is equal to “false”. Therefore, it removes all the outlier records from the data set.



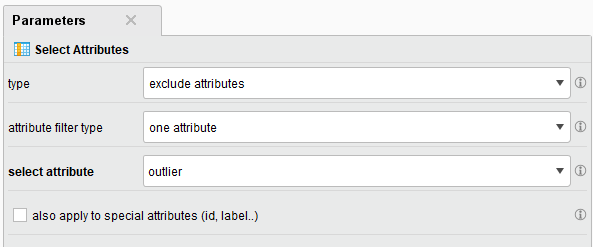


Figure 17, 18: These figures show the usage of the “Select Attributes” operator along with the specified parameters for it. Because all the outliers have been removed in the previous step, having the “outlier” column in the dataset is not useful anymore, so it is being removed.

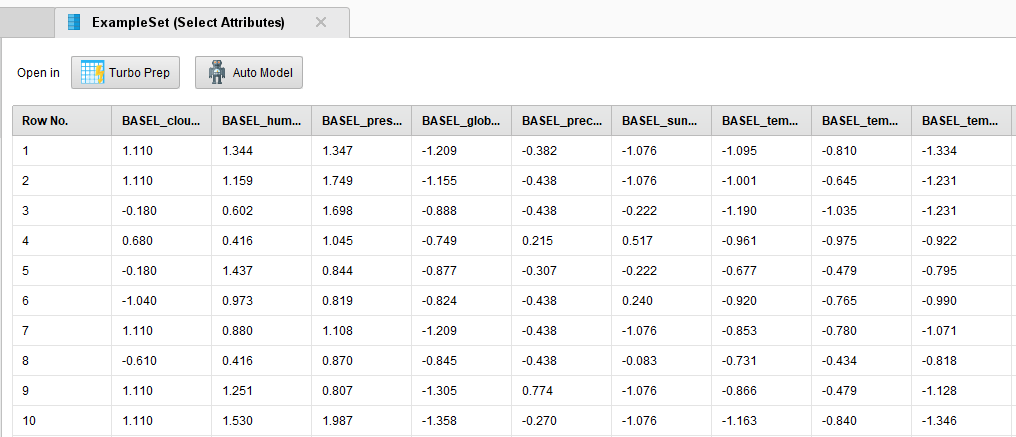


Figure 19: This figure shows the output after completing all the preprocessing stages. The dataset now consists of only numerical values which have been normalized that are free of any duplicates, missing values, and outliers.

PCA has been used as a dimensionality reduction technique as part of the data preprocessing; however, most of the attributes are positively correlated, so when attempting to apply PCA only 1 PC gets generated. Therefore, dimensionality reduction is not valid for this type of dataset.

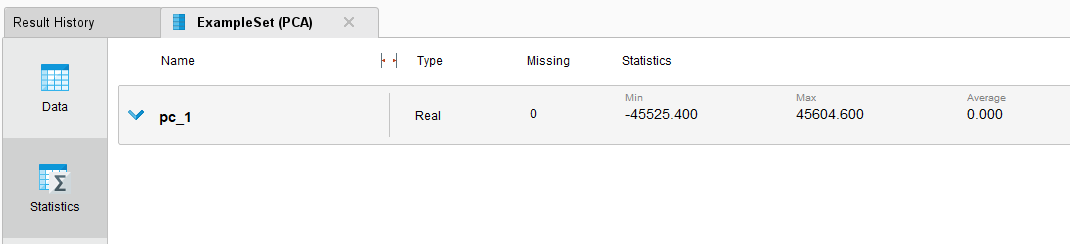


Figure 20: This figure shows the statistics output after attempting to use PCA. It only provides PC 1 because most of the attributes are highly correlated.

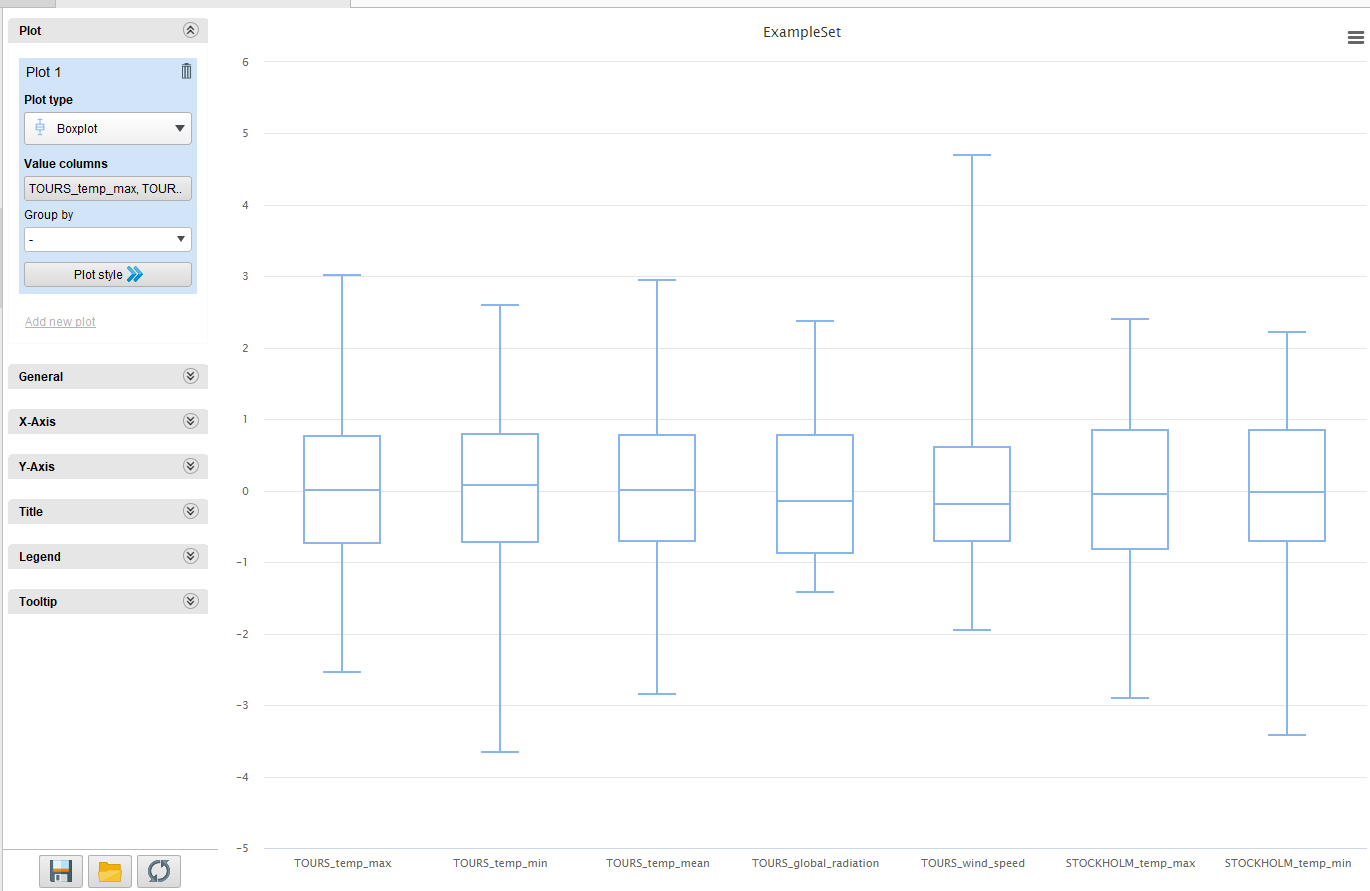


Figure 21: This figure shows the box plot diagram for some of the dataset’s attributes after the preprocessing process. A lot of improvement has occurred for the attributes’ range after the normalization process and removal of outliers. The data is more evenly distributed between positive and negative numbers than the box plot generated before the data preprocessing in Task 2.

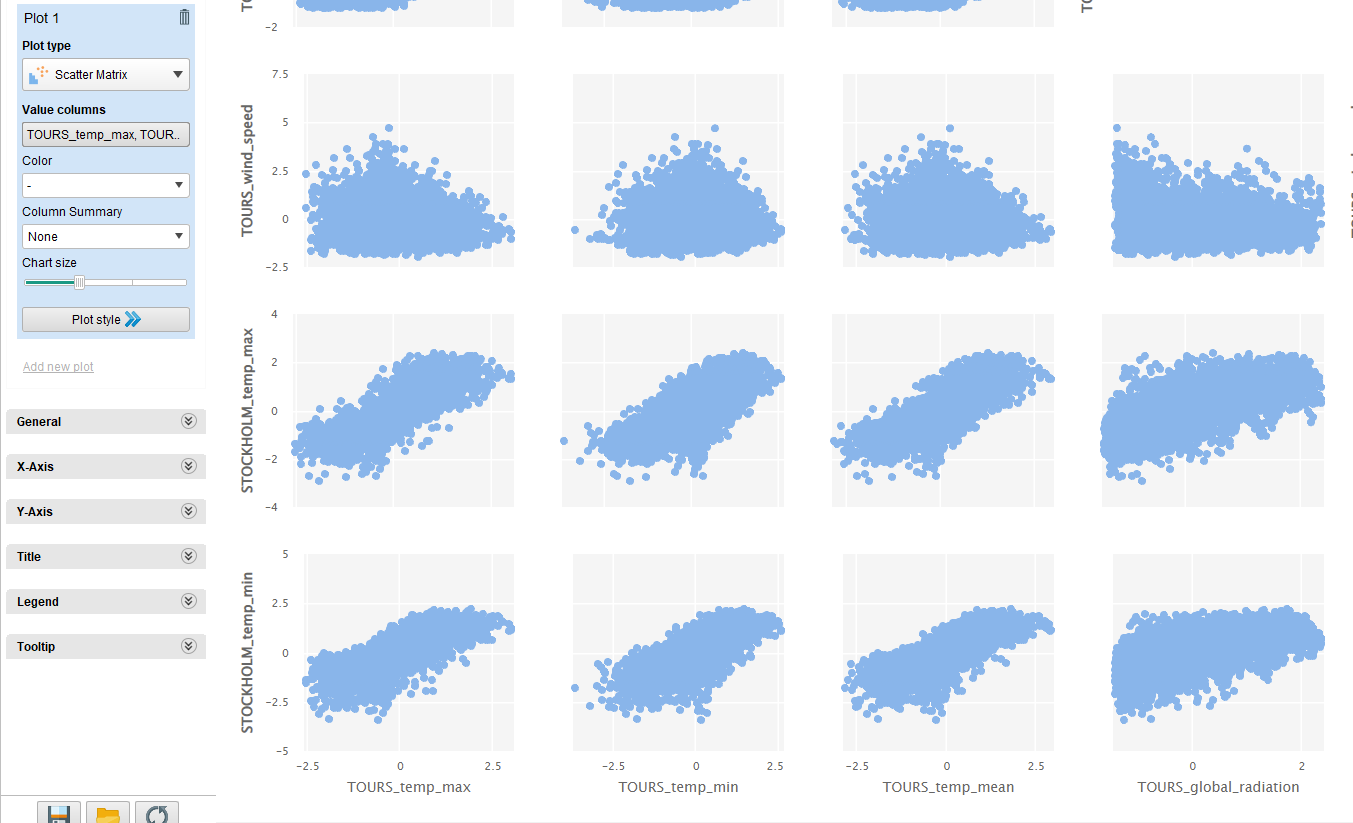
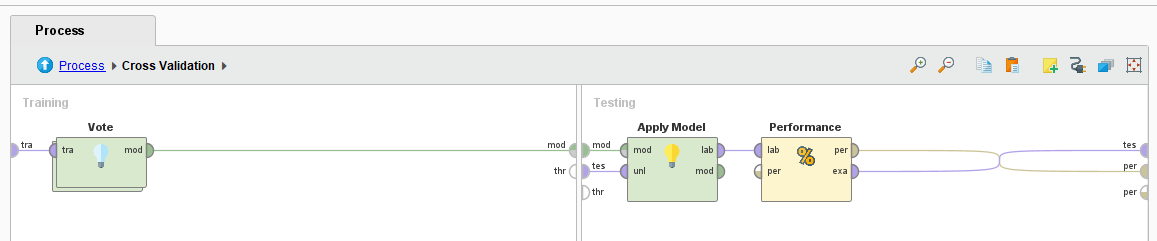


Figure 22: This figure shows the scatter matrix for some of the dataset’s attributes after the preprocessing process. The scatter plots look almost identical to the scatter plots presented before the data preprocessing in Task 2. The only difference is that the ranges have been minimized due to the normalization process. As a result, it can be confirmed that there has been no change to any correlation between the attributes after the data preprocessing.

# Task 4

Supervised learning has been used because the dataset is labeled. For building the data model, rather than dividing the data into a normal split for training and testing sets, cross-validation has been used to obtain more precise results. It trains on one part then tests on the other then takes an iterative approach to make sure all the data can be used for testing, training, and validation. After running multiple tests, the best results have been generated by doing 11 or 14 folds on the dataset.







Figures 23, 24, 25: These figures show the process and the operators used to build the ensemble model by using the “Vote” operator then using different classification techniques.

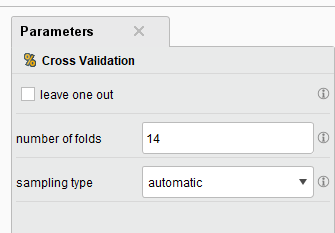


Figure 26: This figure shows the parameter settings for the cross-validation operator. The value 14 was used as it has been proven that it is the most suitable after multiple trials and errors to generate the model with the best outcome.

Three supervised learning techniques have been used to build the classification model.

1. KNN

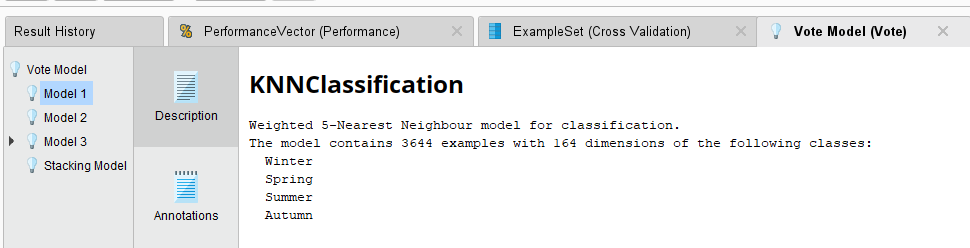
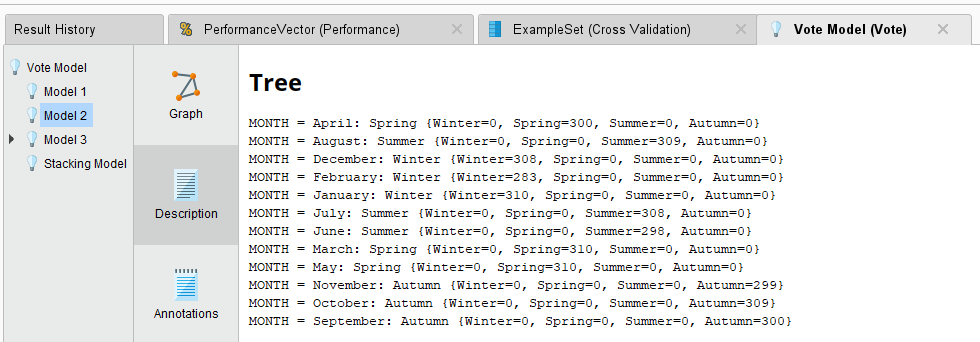
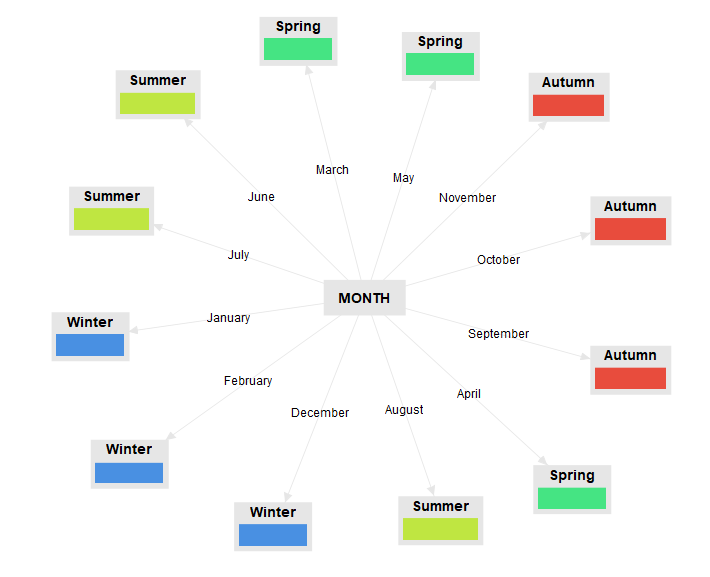


Figure 27: This figure shows the different classifications generated by using KNN.

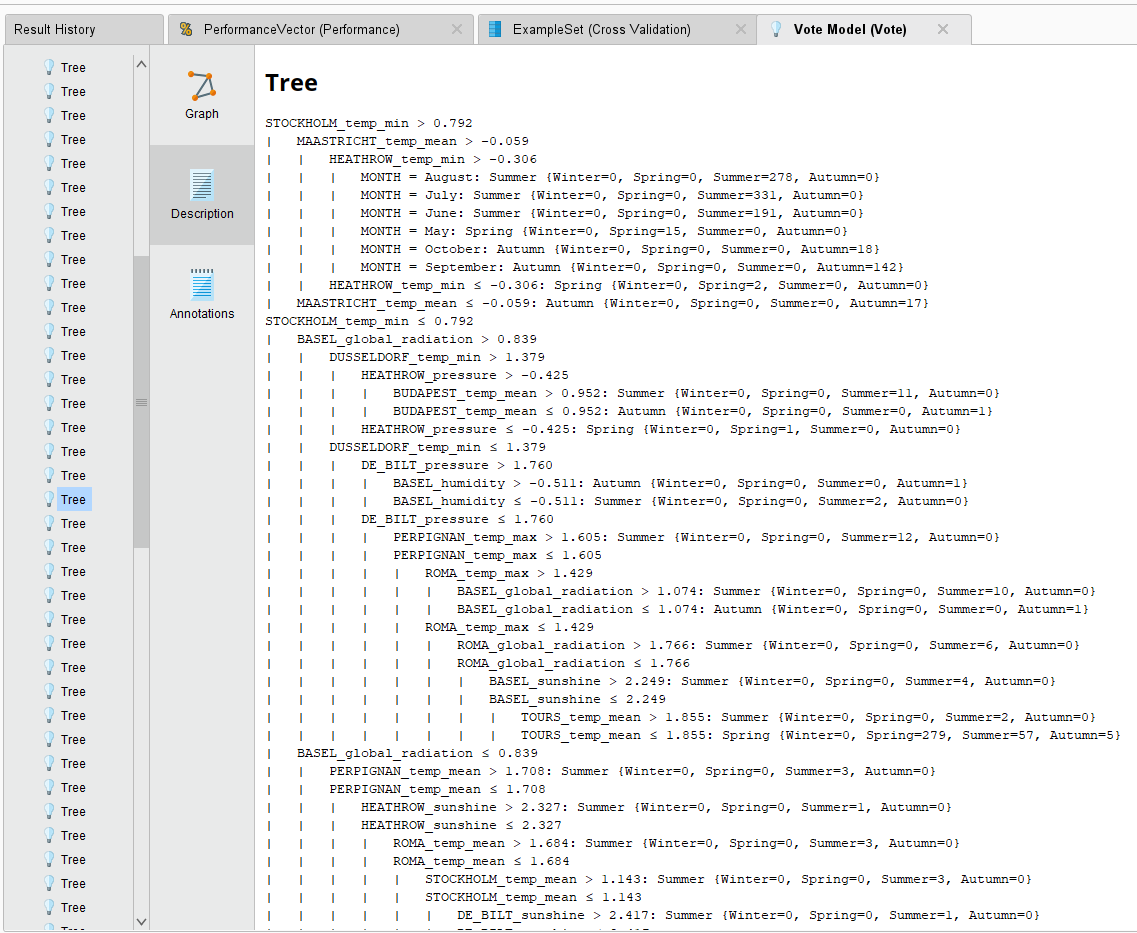
1. Decision Tree

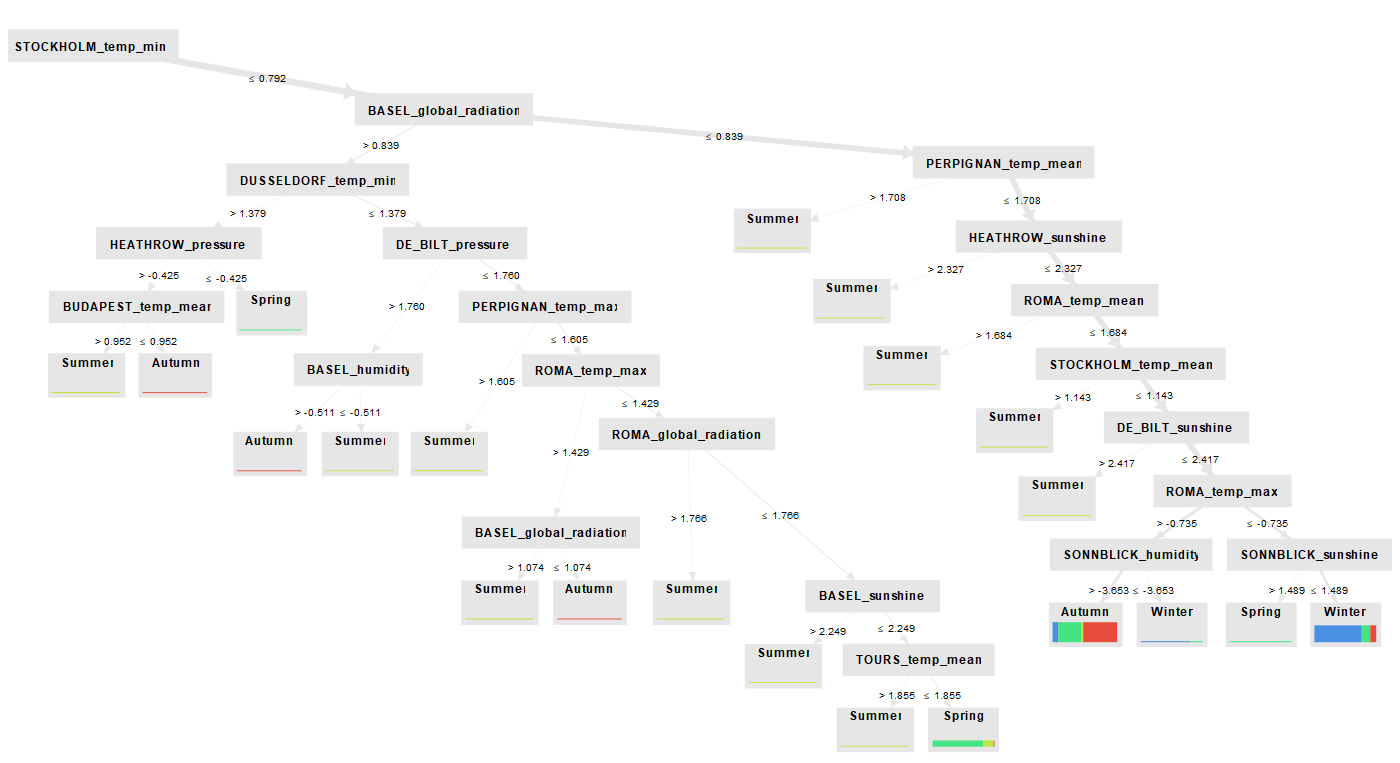




Figures 28, 29: These figures show the decision tree description and the balloon visualization of the tree for making classification decisions.

1. Random Forest





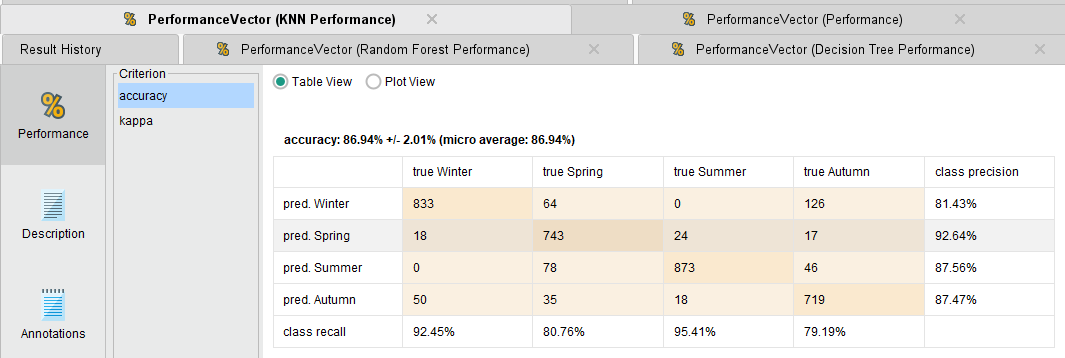
Figures 30, 31: These figures show one of the many trees generated from the random forest classification. The description and visualization of the tree show how each decision has been made based on the attributes’ conditions.

# Task 5

To acquire the best classification technique from the three techniques used, it is better to evaluate the performance of each technique then compare the results.

1. KNN

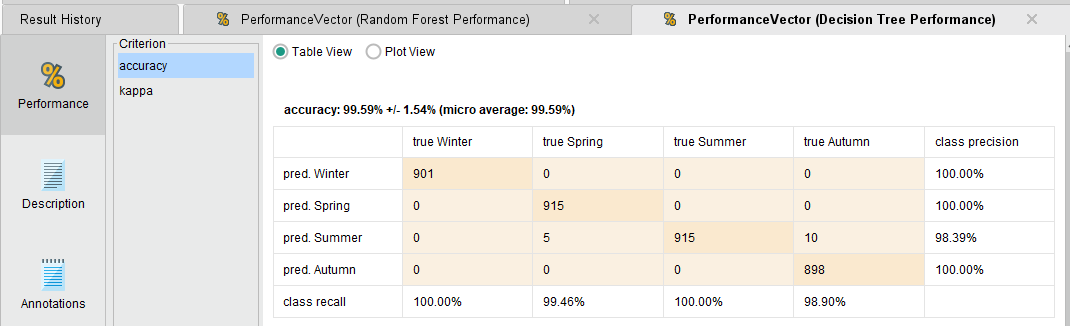
* Accuracy: 86.94%
* Precision:
  + Winter: 81.43%
  + Spring: 92.64%
  + Summer: 87.56%
  + Autumn: 87.47%
* Recall:
  + Winter: 92.45%
  + Spring: 80.76%
  + Summer: 95.41%
  + Autumn: 79.19%

Figure 32: This figure shows the performance of using KNN as a classification technique, showing the confusion matrix.

The confusion matrix demonstrates that the model has performed well in classifying the four seasons, and the high accuracy and kappa value suggests that it is a reliable model.

1. Decision Tree

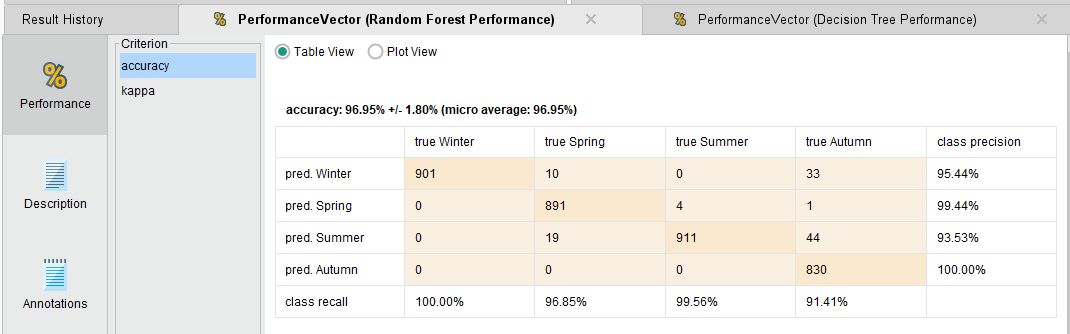
* Accuracy: 99.59%
* Precision:
  + Winter: 100%
  + Spring: 100%
  + Summer: 98.39%
  + Autumn: 100%
* Recall:
  + Winter: 100%
  + Spring: 99.46%
  + Summer: 100%
  + Autumn: 98.9%

Figure 33: This figure shows the performance of using decision tree as a classification technique, showing the confusion matrix.

The model has performed extremely well with the classification process according to the confusion matrix. The very high accuracy and kappa value suggests that it is a highly reliable model for classification.

1. Random Forest

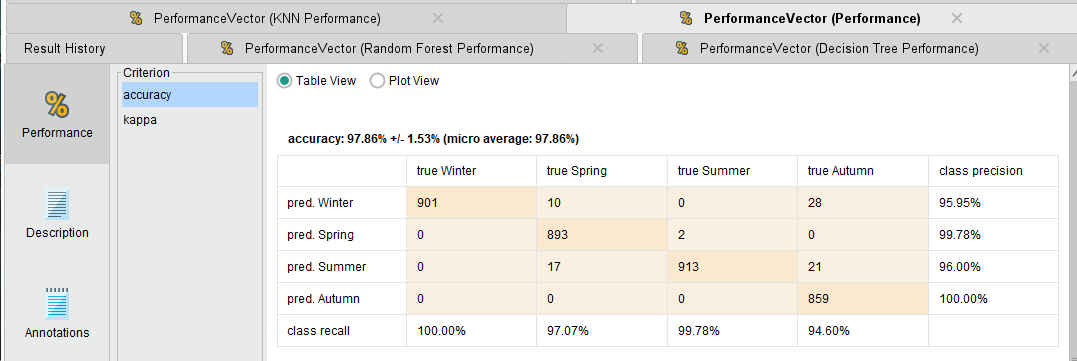
* Accuracy: 96.9%
* Precision:
  + Winter: 95.44%
  + Spring: 99.44%
  + Summer: 93.53%
  + Autumn: 100%
* Recall:
  + Winter: 100%
  + Spring: 98.85%
  + Summer: 99.56%
  + Autumn: 91.41%

Figure 34: This figure shows the performance of using random forest as a classification technique, showing the confusion matrix.

The model has performed well with the classification process. The kappa value of 0.959 indicates substantial agreement between the predicted and actual labels and that it is a highly reliable model for classification.

1. Ensemble Model (Vote)

* Accuracy: 97.86%
* Precision:
  + Winter: 95.95%
  + Spring: 99.78%
  + Summer: 96.00%
  + Autumn: 100%
* Recall:
  + Winter: 100%
  + Spring: 97.07%
  + Summer: 99.78%
  + Autumn: 94.60%

Figure 35: This figure shows the performance of using ensemble model formed by doing a vote between KNN, decision trees, and random forest as classification technique, showing the confusion matrix.

The model correctly classified the majority of the instances with only minimal misclassifications, so it appears to be highly reliable for the classification process.

# Task 6

Overall, the model with the most accurate results has been decision tree. It was able to classify 99.59% of the instances correctly. The predictions were almost flawless and it has proven to be the most effective classification technique out of all the other techniques.  
Random forest and ensemble modeling also had a very good result, classifying the majority of the instances correctly with a very low misclassification rate. Both those classification techniques can be reliable to perform weather forecasting tasks.  
KNN had a good performance as well. Most of the instances were classified correctly. However, compared to how the other models performed, it was the one with the least accuracy and may not be the best choice to go with for weather forecasting tasks.

It can be inferred that data mining techniques, such as decision trees, random forests, and ensemble modeling, can be effective in analyzing large volumes of weather data and finding patterns that can improve the accuracy of weather forecasts. The fact that KNN had a lower accuracy compared to the other models indicates that it may not be the most suitable method for this particular problem.

It is recommended to invest time and money into data mining methods like ensemble modeling, decision trees, and random forests because it can be efficient in evaluating vast amounts of meteorological data and spotting trends that might enhance the precision of weather predictions. Utilizing real-time data can also help produce more accurate short-term weather forecasts. It can help many different divisions. For example, some businesses such as transportation or logistics sectors need quick weather information to make choices, so they may find this to be of great interest.

# Reflection

For our final project in Data Mining, we used data mining techniques in Rapidminer to explore weather forecasting. By collecting and processing large volumes of weather data, we aimed to uncover patterns and relationships to improve the accuracy of predictions. Our project represents an innovative application of data mining that has the potential to improve weather forecasting for various applications. We faced many challenges during this project but fortunately we managed to overcome them.

In the initial phase of our project, we were tasked with articulating the business objective of our project, as defining our business goal at the start of the project was a key step because it gave us a clear idea of what we wanted to achieve and how it could benefit us. Then we had to choose a good dataset related to our topic, selecting a suitable dataset related to our topic took time due to various issues, including irrelevance, size, and quality. Nonetheless, we were able to identify a good dataset after careful consideration and evaluation. After preprocessing our dataset, we attempted PCA for dimensionality reduction, but the high positive correlations generated only one principal component. Therefore, we adapted our analysis to utilize the one component and derive insights. Another challenge we faced was to determine the best number of folds to be done for cross-validation. Moreover, the best choice for us was trial and error. This technique allowed us to get the best result regarding the number of folds to be done for our classification.

In the process of working on our data mining project, we were exposed to several innovative techniques for solving complicated issues in a more effective way. We advanced considerably and accomplished considerable gains in our analytical ability by utilizing these strategies. As the project ends, we are happy to say that we now have high confidence as data mining practitioners. We are confident that the knowledge we have obtained from working on this project will be essential in our future undertakings and allow us to significantly advance the field of data mining.

# Roles & Logs

