ANTALYA BİLİM ÜNİVERSİTESİ

CS 447
Introduction to Data Science

Spring 2020

Instructor: *Dr. Şadi Evren Şeker*

Project Presentation

Submitted by
190201038 - Ibrahim Berber

Project Idea: *Rain Prediction in Australia*

Approaches

- Random Forest Classifier
- Logistic Regression

Tools and Platforms

ANACONDA.

Open for Innovation ®
KNIME

1. Introduction

- Weather prediction and data science
- Rain in Australia Dataset
- Classification: raining or not raining
- Daily weather observations
- Data is prepared by Australian weather stations

Dataset and Data Availability: What is the data?

- File size 13 MB
- Target variable: **`RainTomorrow`**
- Available at Kaggle. (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.)

```
In [1]: import pandas as pd

        weather_dataframe = pd.read_csv('weatherAUS.csv')

In [2]: weather_dataframe.head()
```

Out[2]:

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...
0	12/1/2008	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...
1	12/2/2008	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...
2	12/3/2008	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...
3	12/4/2008	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...
4	12/5/2008	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...

5 rows × 24 columns

Figure: Overview of data

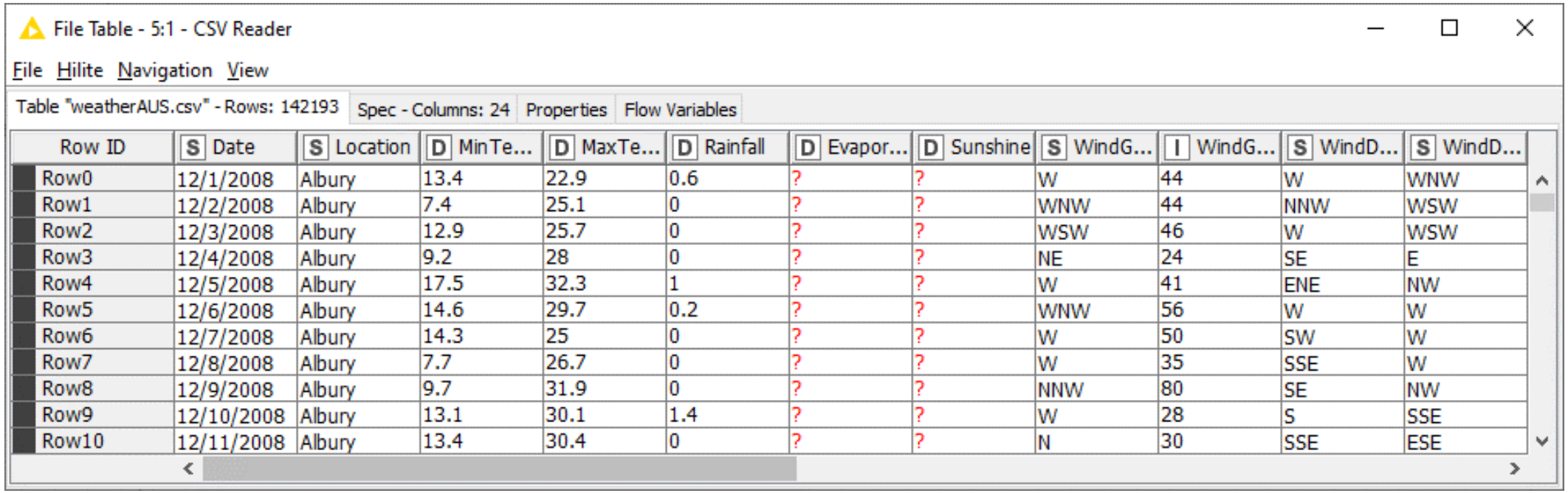
Dataset: What are the columns?

- Column types: Date, String, Integer, Decimal
- Data types: Categorical, Continuous, Numerical

Columns	Description
Date	The date of observation
Location	The common name of the location of the weather station
MinTemp	The minimum temperature in degrees Celsius
Rainfall	The amount of rainfall recorded for the day in mm
Sunshine	The number of hours of bright sunshine in the day.
RainTomorrow	The target variable. Did it rain tomorrow?
…	…

Table: Column Descriptions

Importing in KNIME: CSV Reader ( )

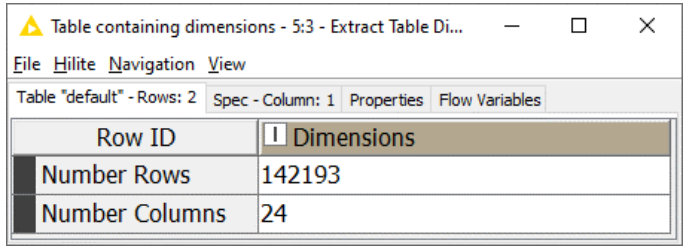


Row ID	S Date	S Location	D MinTe...	D MaxTe...	D Rainfall	D Evapor...	D Sunshine	S WindG...	I WindG...	S WindD...	S WindD...
Row0	12/1/2008	Albury	13.4	22.9	0.6	?	?	W	44	W	WNW
Row1	12/2/2008	Albury	7.4	25.1	0	?	?	WNW	44	NNW	WSW
Row2	12/3/2008	Albury	12.9	25.7	0	?	?	WSW	46	W	WSW
Row3	12/4/2008	Albury	9.2	28	0	?	?	NE	24	SE	E
Row4	12/5/2008	Albury	17.5	32.3	1	?	?	W	41	ENE	NW
Row5	12/6/2008	Albury	14.6	29.7	0.2	?	?	WNW	56	W	W
Row6	12/7/2008	Albury	14.3	25	0	?	?	W	50	SW	W
Row7	12/8/2008	Albury	7.7	26.7	0	?	?	W	35	SSE	W
Row8	12/9/2008	Albury	9.7	31.9	0	?	?	NNW	80	SE	NW
Row9	12/10/2008	Albury	13.1	30.1	1.4	?	?	W	28	S	SSE
Row10	12/11/2008	Albury	13.4	30.4	0	?	?	N	30	SSE	ESE

Figure: Data is imported in KNIME

Data Specs:

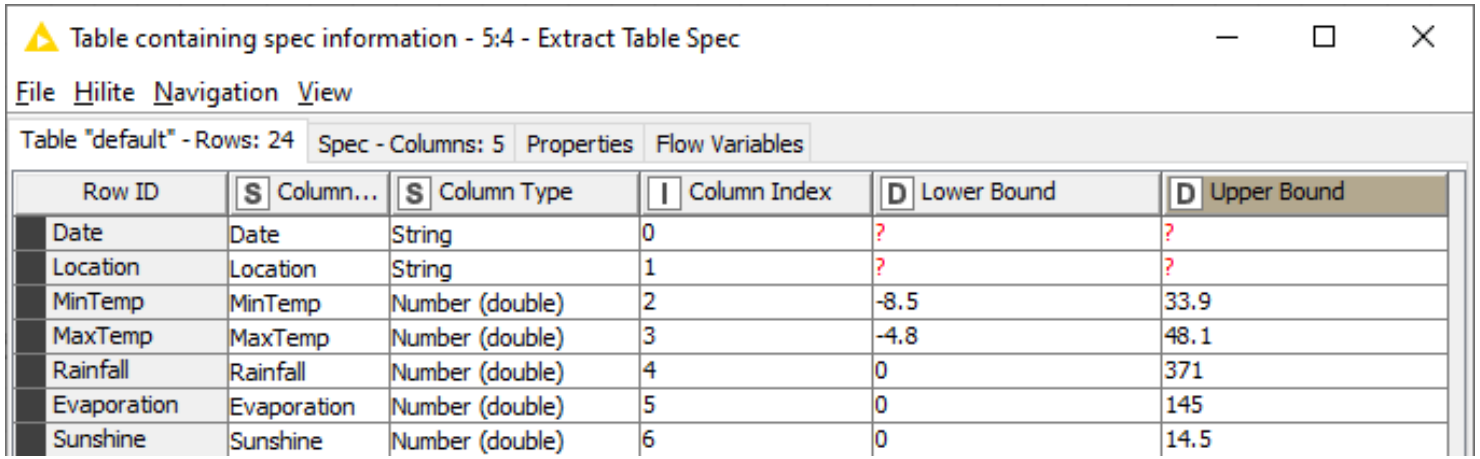
Extract Table Dimension



Row ID	Dimensions
Number Rows	142193
Number Columns	24

Figure: Data dimensions

Extract Table Specs



Row ID	Column...	Column Type	Column Index	Lower Bound	Upper Bound
Date	Date	String	0	?	?
Location	Location	String	1	?	?
MinTemp	MinTemp	Number (double)	2	-8.5	33.9
MaxTemp	MaxTemp	Number (double)	3	-4.8	48.1
Rainfall	Rainfall	Number (double)	4	0	371
Evaporation	Evaporation	Number (double)	5	0	145
Sunshine	Sunshine	Number (double)	6	0	14.5

Figure: Column info

Data Preprocessing

- Excluding columns: Too many missing values (Sunshine, Evaporation, Cloud3pm and Cloud9am)
- Excluding columns: Purpose and requirement (Location, RISK_MM, Date*)

Statistics Table - 2:4 - Statistics

File Hilite Navigation View

Table "default" - Rows: 17 | Spec - Columns: 16 | Properties | Flow Variables

Row ID	Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Overall ...	No. missings
Sunshine	Sunshine	0	14.5	7.625	3.782	14.3	-0.503	-0.82	567,113.7	67816
Evaporation	Evaporation	0	145	5.47	4.189	17.544	3.747	45.068	444,970.2	60843
Cloud3pm	Cloud3pm	0	9	4.503	2.721	7.402	-0.224	-1.458	383,215	57094
Cloud9am	Cloud9am	0	9	4.437	2.887	8.335	-0.224	-1.541	392,851	53657
Pressure9am	Pressure9am	980.5	1,041	1,017.654	7.105	50.488	-0.096	0.236	130,441,841.	14014
Pressure3pm	Pressure3pm	977.1	1,039.6	1,015.258	7.037	49.515	-0.046	0.133	130,168,28...	13981
WindGustSpeed	WindGustSp...	6	135	39.984	13.589	184.656	0.874	1.418	5,314,832	9270
Humidity3pm	Humidity3pm	0	100	51.483	20.798	432.547	0.035	-0.511	7,134,614	3610
Temp3pm	Temp3pm	-5.4	46.7	21.687	6.938	48.13	0.24	-0.146	3,024,653.6	2726

Figure: Number of missing values for columns

Data Preprocessing

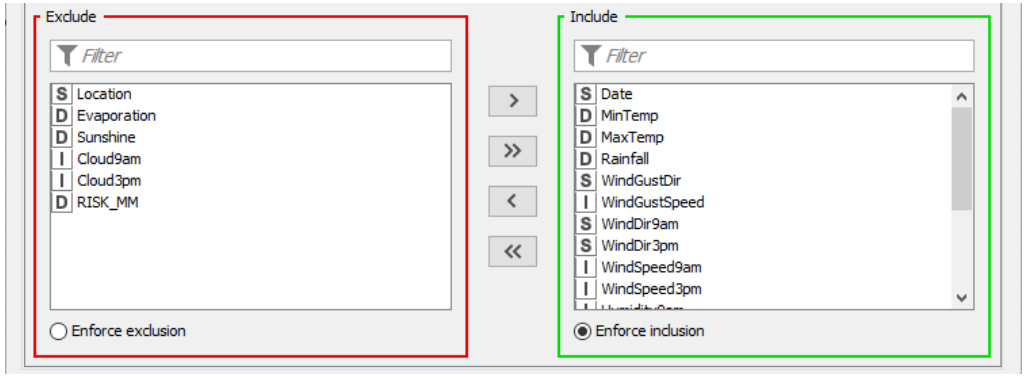


Figure: Filtering*

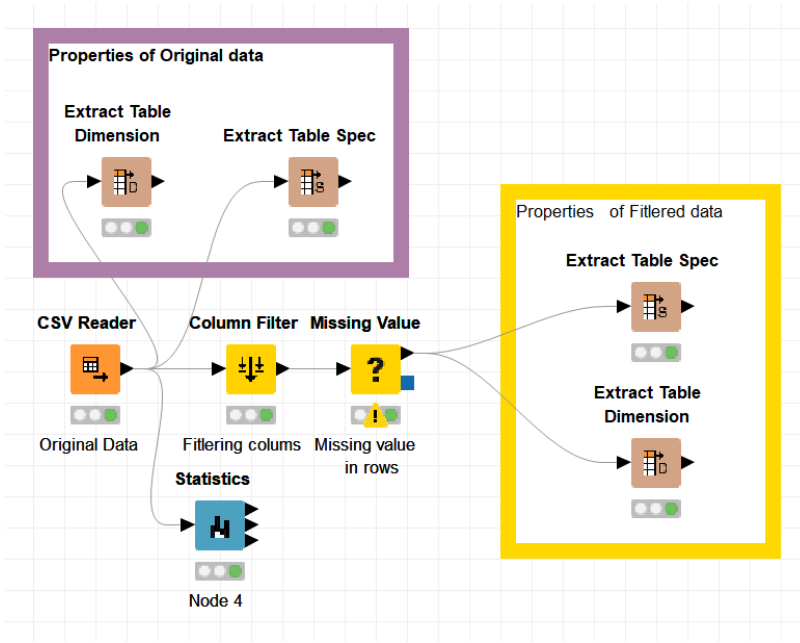


Figure: Nodes in KNIME

2. Analysis of Data

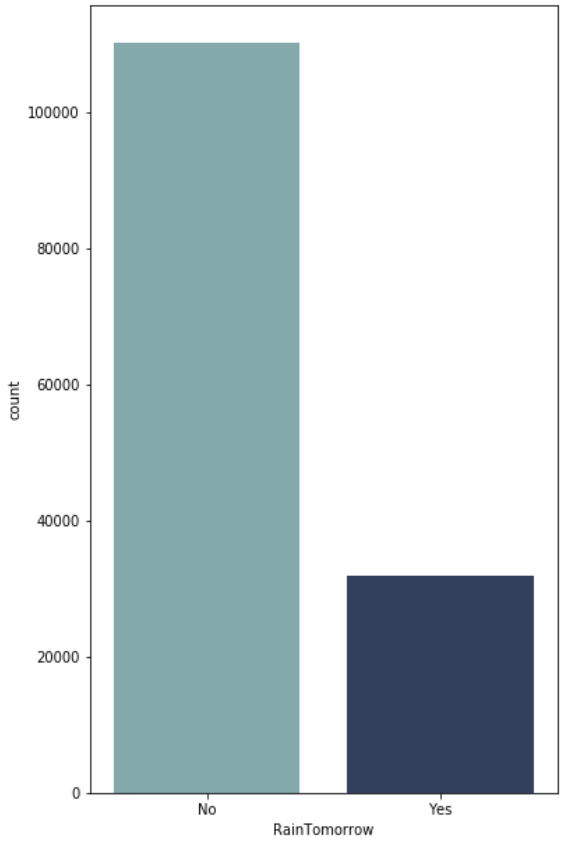
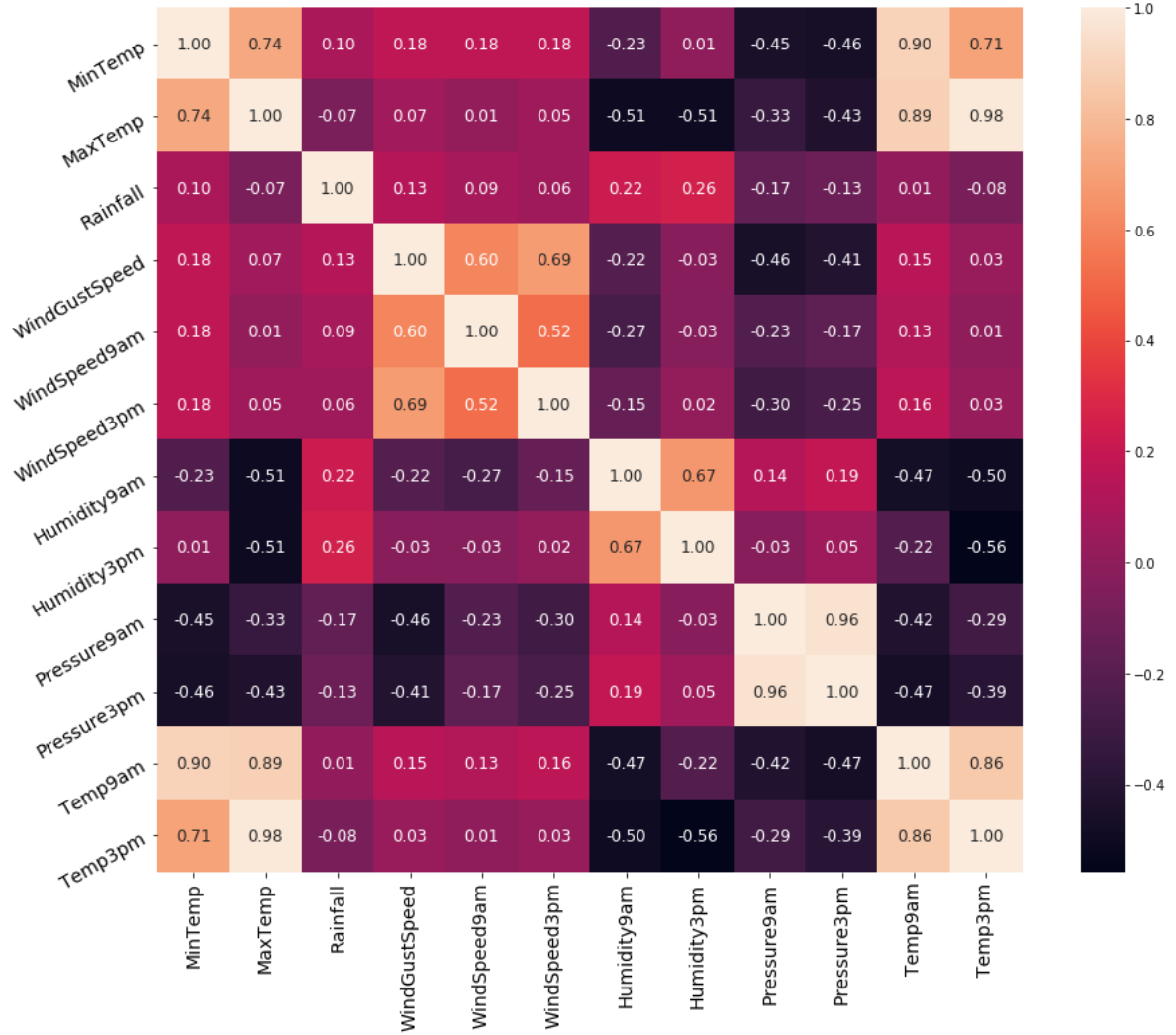


Figure: # of Y/N

Row ID	count
No	110316
Yes	31877

Figure: Occurrences

Figure: Heatmap Correlation



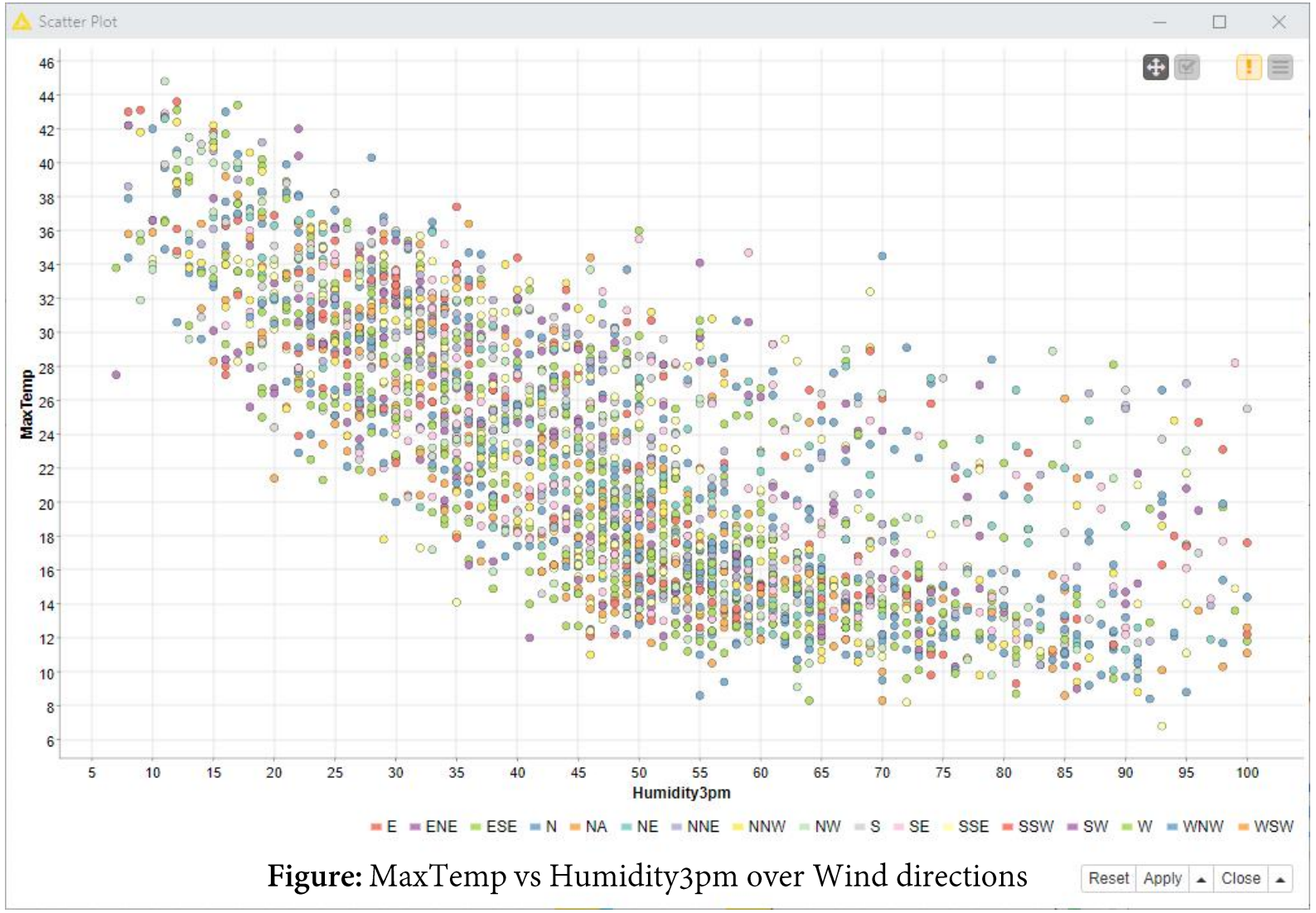
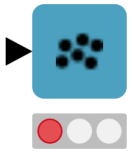


Figure: MaxTemp vs Humidity3pm over Wind directions

Visualization Example: Scatter Plots in KNIME



Scatter Plot

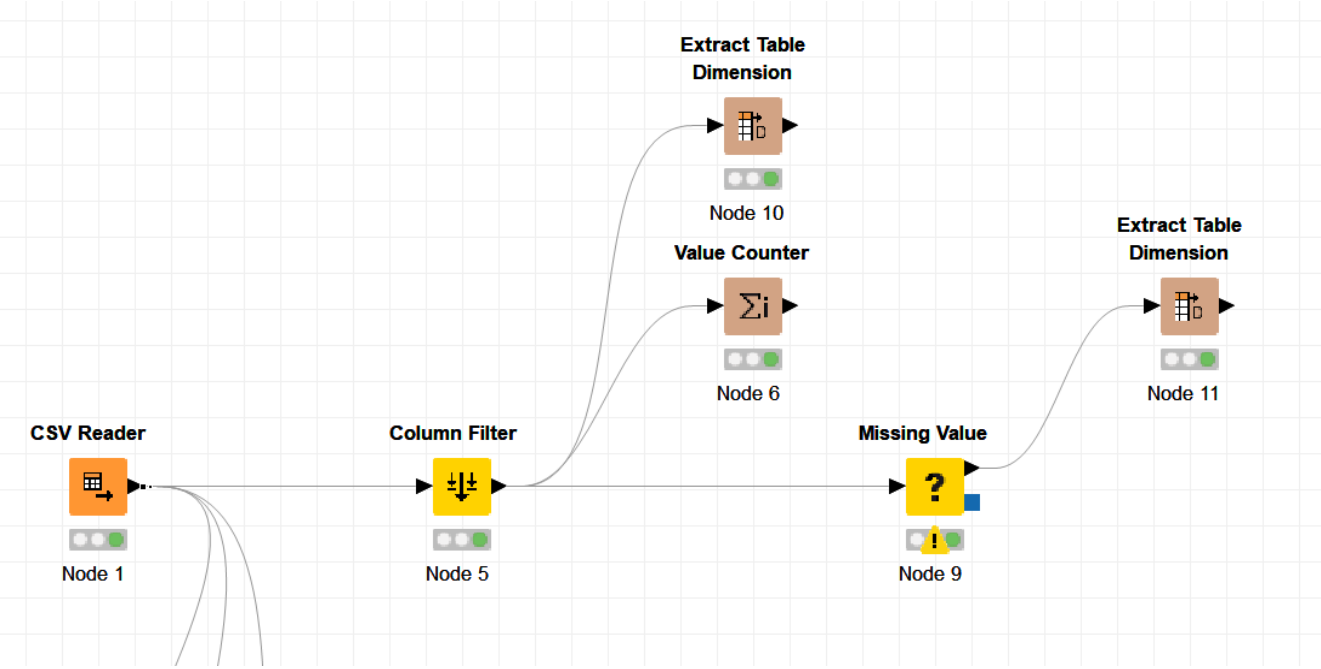


Figure: Related nodes

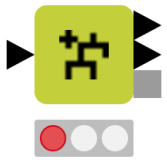
Operation	# of Rows	# of Columns
Original (No operation applied)	142,193	24
Filtered (7 columns excluded)	142,193	17
Missing Value node	119590	17

Table: Dimensions of Dataset

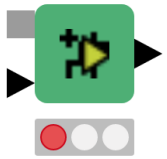
3. Implementation

Approaches

- Random Forest Classifier



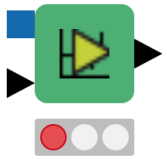
Random Forest Learner



Random Forest Predictor

- Logistic Regression

Logistic Regression Learner



Logistic Regressing Predictor

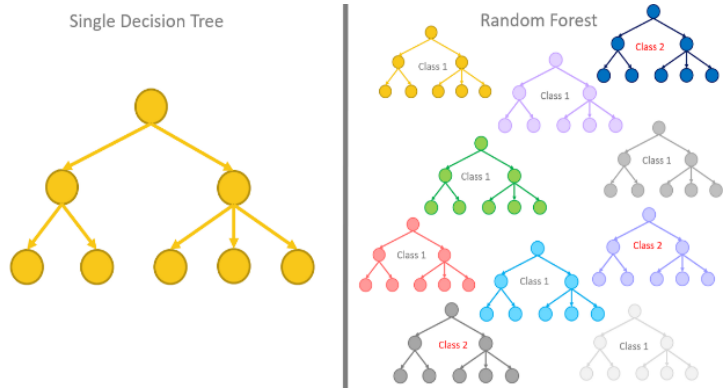


Figure: Random Forest

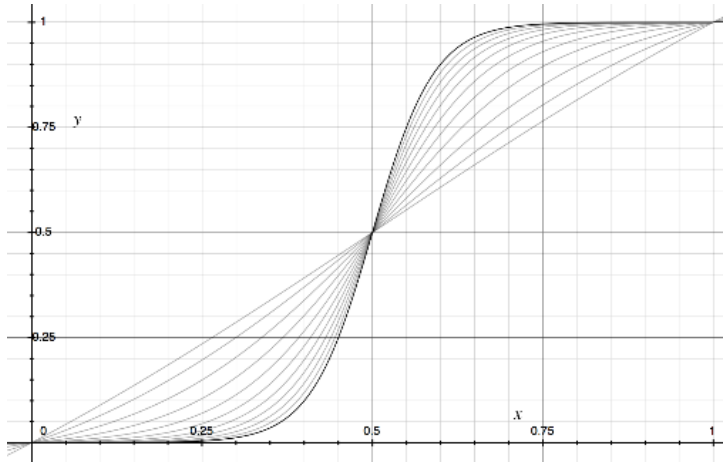


Figure: Logistic Regression

Approach I: Random Forest

- Confusion matrix (CM)
- Partitioning:
  - none, 60, 70 and 80 %
  - ± 7% difference
- Overfitting, Impact of partitioning

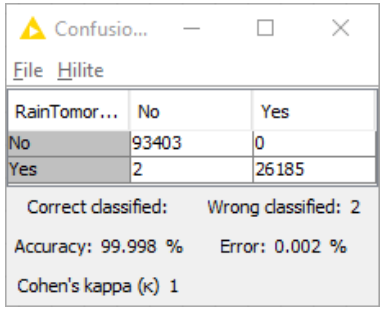


Figure: CM with no partition

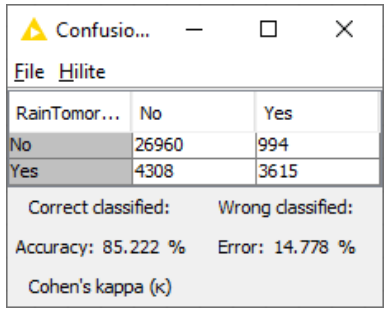


Figure: CM with partitioning

Attribute Statistics - 0:16 - Random Forest Learner

File Hilite Navigation View

Table "Tree Ensemble Column Statistic" - Rows: 16 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (level 1)	#candidates (level 2)
MinTemp	0	7	8	21	41	101
MaxTemp	1	0	10	22	47	105
Rainfall	17	21	31	27	51	95
WindGustDir	1	15	45	31	53	102
WindGustSpeed	9	30	49	29	63	99
WindDir9am	0	4	13	25	52	95
WindDir3pm	0	2	21	23	50	117
WindSpeed9am	0	1	7	24	63	110
WindSpeed3pm	1	0	6	20	55	108
Humidity9am	15	18	25	33	44	97
Humidity3pm	19	51	71	19	56	84
Pressure9am	4	8	27	25	51	89
Pressure3pm	7	12	38	27	34	104
Temp9am	0	3	6	22	35	102
Temp3pm	3	7	18	27	49	103
RainToday	23	21	17	25	56	89

Table: Accuracies on different partitioning

Approach I: Random Forest: ROC Curves ( )

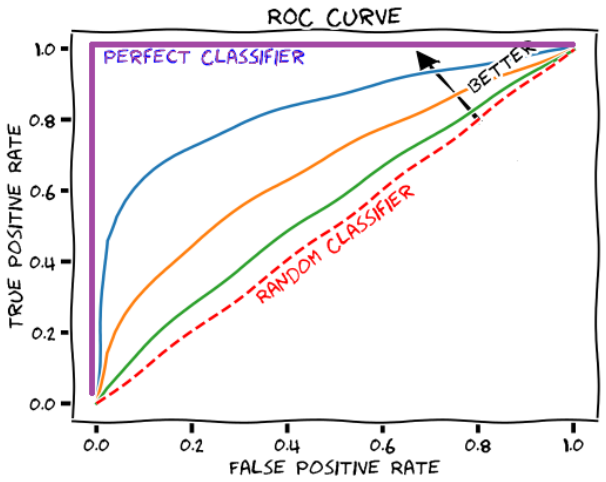


Figure: ROC curve explained

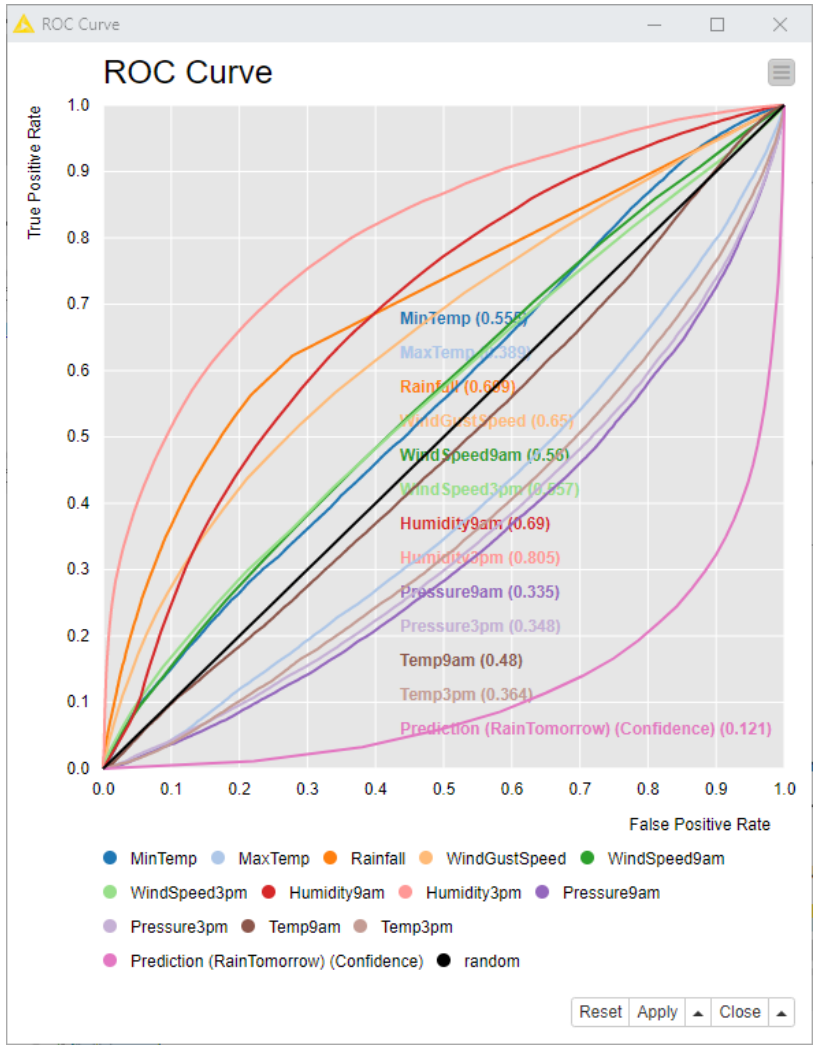


Figure: ROC curve with no partitioning

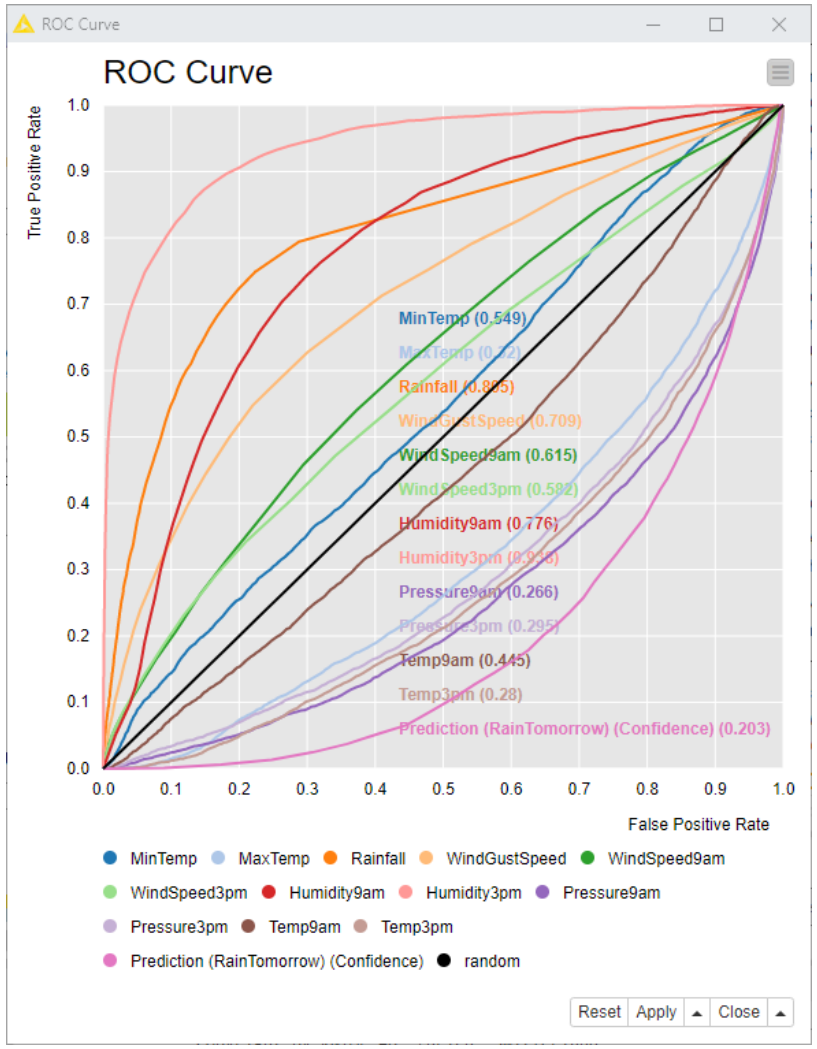


Figure: ROC curve with partitioning of 70 %

Approach I: Random Forest: Parameter selection & Accuracies

- Normalization
  - None, *Z-Score*

Normalizer

- Limiting parameters:
  - *Tree depth*: 2, 5, 10, unlimited
  - *Min node size*: 1, 2, 5, 20, not specified
- Number of model: 2, 20, 100, 500
- Consistent accuracy with 20+ models
- Normalization not affect accuracy

Tree Options	Tree depth	Min Node size	# of Model	Accuracy
Information Gain	2	1	2	79.463 %
Information Gain	5	2	20	83.747 %
Information Gain	10	2	100	85.071 %
Information Gain	Not limited	Not specified	2	81,771 %
Information Gain	Not limited	Not specified	100	85.236 %
Information Gain	Not limited	Not specified	500	85.328 %
Information Gain Ratio	2	1	2	79.315 %
Information Gain Ratio	5	2	20	83.354 %
Information Gain Ratio	10	2	100	84.756 %
Information Gain Ratio	Not limited	Not specified	2	82.069 %
Information Gain Ratio	Not limited	Not specified	100	85.222 %
Information Gain Ratio	Not limited	Not specified	500	85.358 %
Gini Index	2	1	2	79.616 %
Gini Index	5	2	20	83.987 %
Gini Index	10	2	100	84.993 %
Gini Index	Not limited	Not specified	2	81.520 %
Gini Index	Not limited	Not specified	100	85.135 %
Gini Index	Not limited	Not specified	500	85.303 %

Table: Random Forest Accuracies

Approach I: Random Forest:
What is the best parameters & settings?

- Works well on most settings
- *Tree depth*: 5+
- *Min node size*: 2+
- *Number of model*: 5+
- Consistent accuracy of 85%.

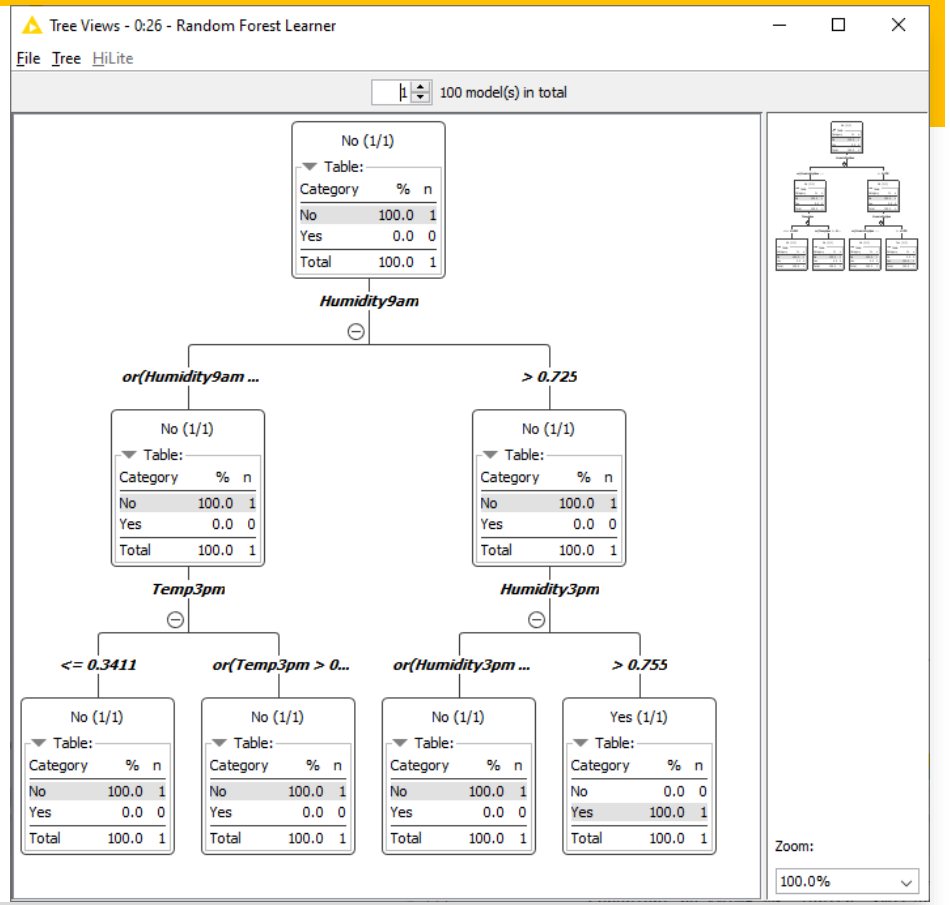


Figure: Example of one of tree with two levels

Random Forest in KNIME

- Partitioning
- Default settings

- Partitioning
- Normalization
- Various parameters

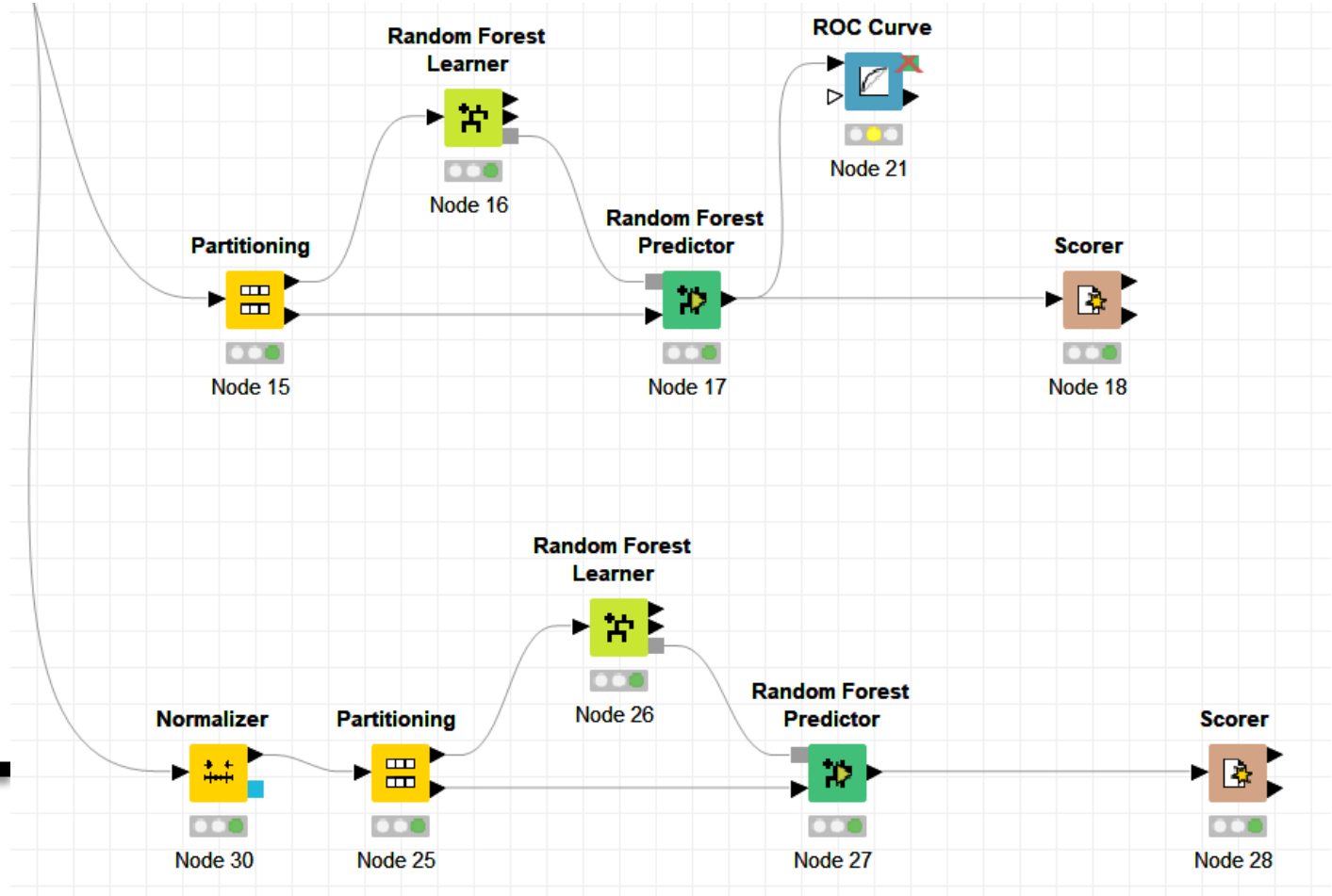


Table: KNIME workflow of Random Forest

Approach II: Logistic Regression

- Confusion matrix (CM)
- Partitioning:
  - none, 60, 70, 80 and 90%
  - Strange accuracies
- Execution with default settings
- Impact of partitioning

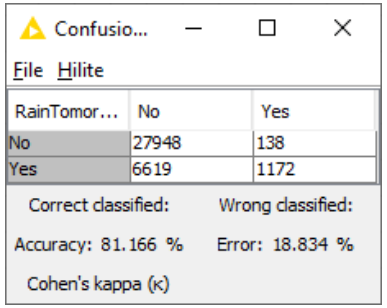


Figure: CM of 70% partitioning

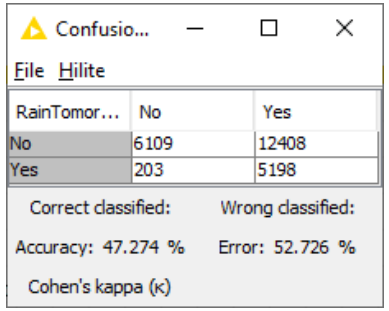


Figure: CM of 80% partitioning

Partitioning	Execution 1	Execution 2	Execution 3	Avg Accuracy
None	82.718 %	72.678 %	60.597 %	71.998 %
60%	84.196 %	83.175 %	83.552 %	83.641 %
70%	81.166 %	84.485 %	83.000 %	82.883 %
80%	47.274 %	83.895 %	83.744 %	71.638 % (err)
90%	77.147 %	78.803 %	81.704 %	79.218 %

Table: Accuracies on different partitioning

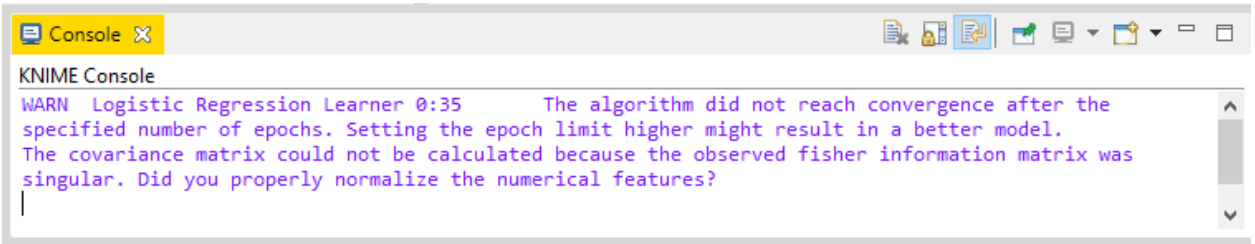


Figure: KNIME Warning

Approach II: Logistic Regression: ROC Curves ( )

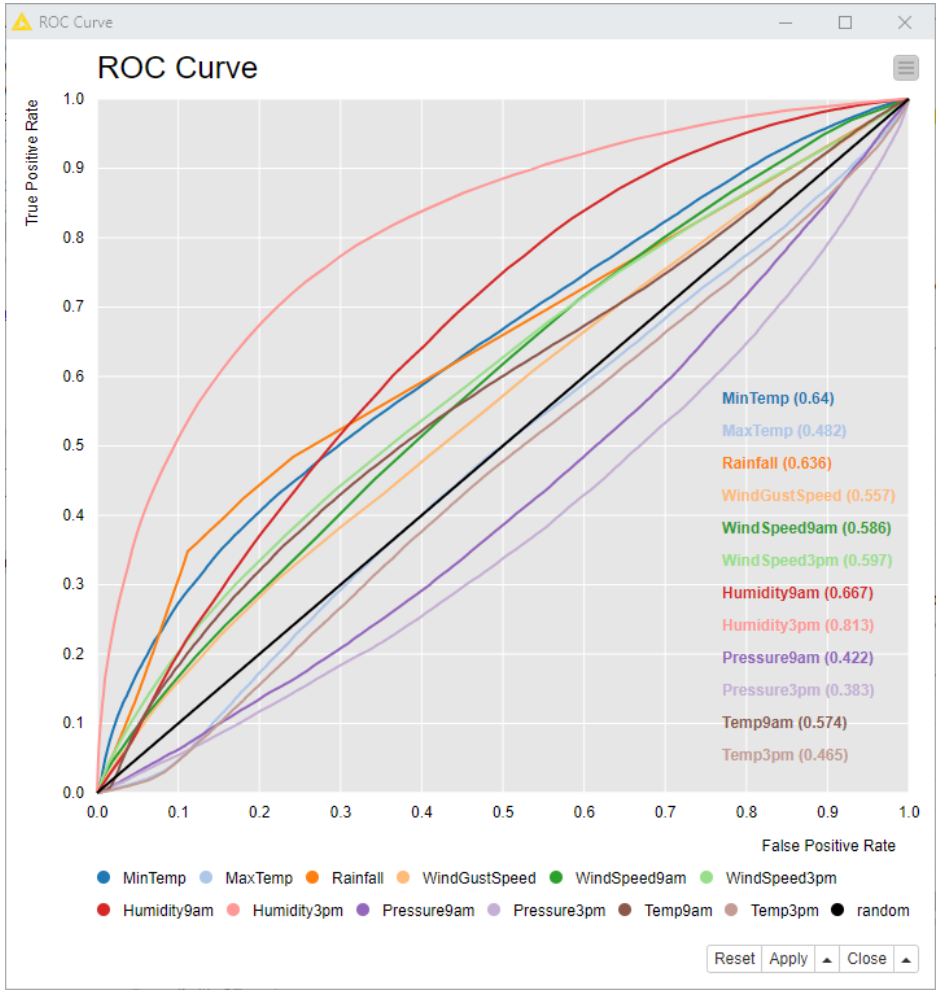


Figure: ROC curve with no partitioning

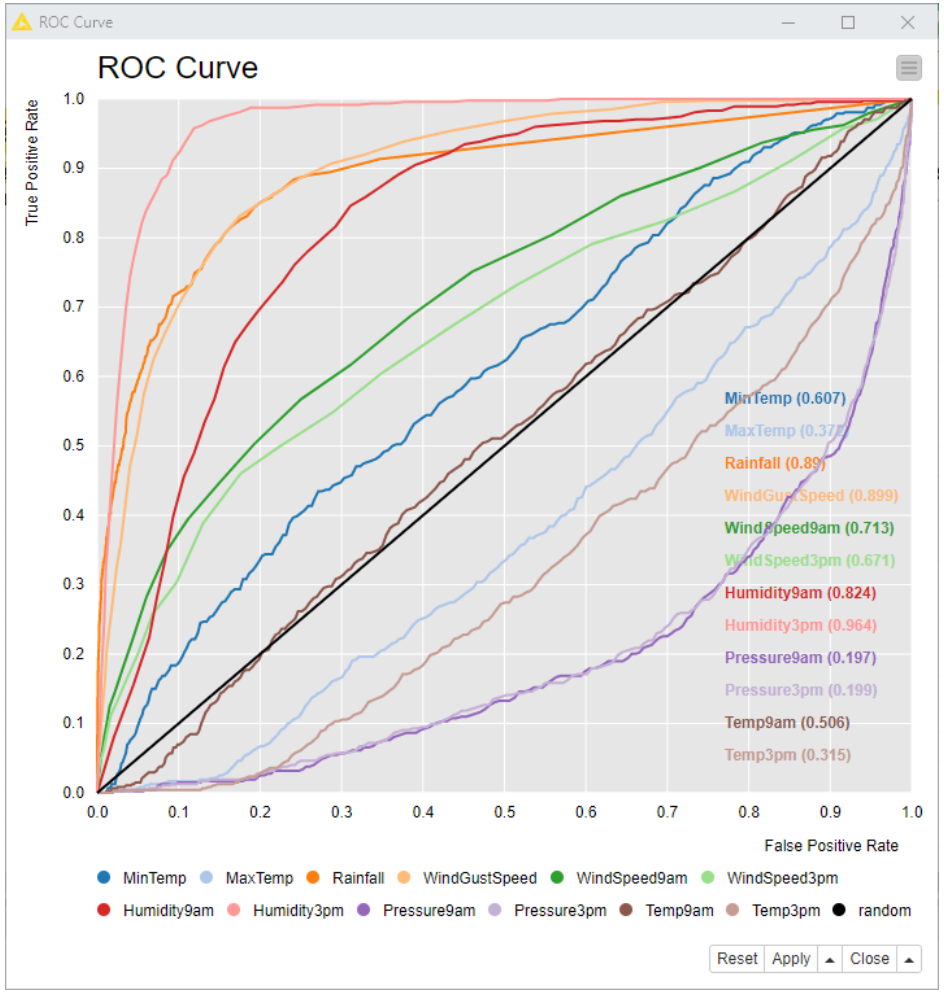


Figure: ROC curve with partitioning of 70 %

Approach II: Logistic Regression: Parameter selection & Accuracies

- Normalization
  - None, *Min-Max*, *Z-Score*

Normalizer

- Regularization:
  - *Uniform*, *Gauss*, *Laplace*
- Parameters:
  - *Number of Epochs*: 100, 500
  - *Learning Rate*: 0.01, 0.10
- Consistent accuracy with normalization
- Normalization is crucial

Normalization	# of Epochs	Learning Rate	Regularization	Accuracy 1	Accuracy 2	Accuracy 3
Not applied	100	Fixed: 0.01	Uniform	78 %	80 %	84 %
Not applied	100	Fixed: 0.01	Gauss	84 %	65 %	79 %
Not applied	100	Fixed: 0.01	Laplace	83 %	84 %	83 %
Not applied	100	Fixed: 0.10	Uniform	84 %	70 %	81 %
Not applied	100	Fixed: 0.10	Gauss	77 %	83 %	75 %
Not applied	100	Fixed: 0.10	Laplace	82 %	73 %	80 %
Not applied	500	Fixed: 0.01	Uniform	81 %	82 %	82 %
Not applied	500	Fixed: 0.01	Gauss	83 %	78 %	81 %
Not applied	500	Fixed: 0.01	Laplace	78 %	84 %	81 %
Not applied	500	Fixed: 0.10	Uniform	78 %	85 %	45 %
Not applied	500	Fixed: 0.10	Gauss	52 %	79 %	73 %
Not applied	500	Fixed: 0.10	Laplace	81 %	82 %	78 %
Min-Max	100	Fixed: 0.01	Uniform	85 %	85 %	85 %
Min-Max	100	Fixed: 0.01	Gauss	85 %	85 %	84 %
Min-Max	100	Fixed: 0.01	Laplace	85 %	85 %	85 %
Min-Max	100	Fixed: 0.10	Uniform	85 %	85 %	85 %
Min-Max	100	Fixed: 0.10	Gauss	84 %	85 %	85 %
Min-Max	100	Fixed: 0.10	Laplace	85 %	85 %	85 %
Min-Max	500	Fixed: 0.01	Uniform	85 %	85 %	85 %
Min-Max	500	Fixed: 0.01	Gauss	85 %	84 %	85 %
Min-Max	500	Fixed: 0.01	Laplace	85 %	85 %	85 %
Min-Max	500	Fixed: 0.10	Uniform	85 %	85 %	85 %
Min-Max	500	Fixed: 0.10	Gauss	84 %	85 %	85 %
Min-Max	500	Fixed: 0.10	Laplace	85 %	85 %	85 %
Z-Score	100	Fixed: 0.01	Uniform	85 %	85 %	85 %
Z-Score	100	Fixed: 0.01	Gauss	85 %	85 %	84 %
Z-Score	100	Fixed: 0.01	Laplace	85 %	84 %	85 %
Z-Score	100	Fixed: 0.10	Uniform	85 %	85 %	85 %
Z-Score	100	Fixed: 0.10	Gauss	85 %	85 %	84 %
Z-Score	100	Fixed: 0.10	Laplace	84 %	85 %	85 %
Z-Score	500	Fixed: 0.01	Uniform	85 %	85 %	85 %
Z-Score	500	Fixed: 0.01	Gauss	85 %	85 %	85 %
Z-Score	500	Fixed: 0.01	Laplace	85 %	84 %	85 %
Z-Score	500	Fixed: 0.10	Uniform	85 %	85 %	84 %
Z-Score	500	Fixed: 0.10	Gauss	84 %	85 %	85 %
Z-Score	500	Fixed: 0.10	Laplace	85 %	84 %	84 %

Table: Logistic Regression Accuracies

Approach II: Logistic Regression: What is the best parameters & settings?

- *Number of epochs:* 5000, 10K and 15K.
- *Epsilon*: 1.0E-6 and 1.0E-7
- *Learning rate*: LineSearch and fixed of values of 1.0E-3, 1.0E-4
- *Regularization*: Laplace and Gauss
- Consistent accuracy of 85%.

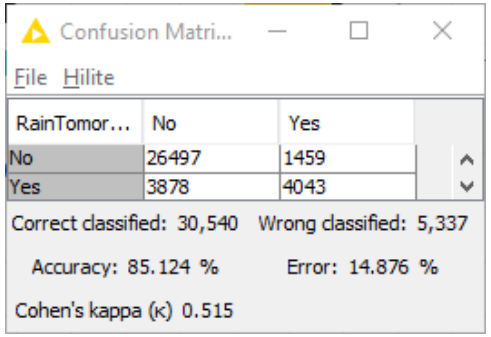


Figure: CM of extreme parameters

Logistic Regression in KNIME

- Default parameters
- No partitioning

- Partitioning

- Partitioning
- Normalization
- Regularization

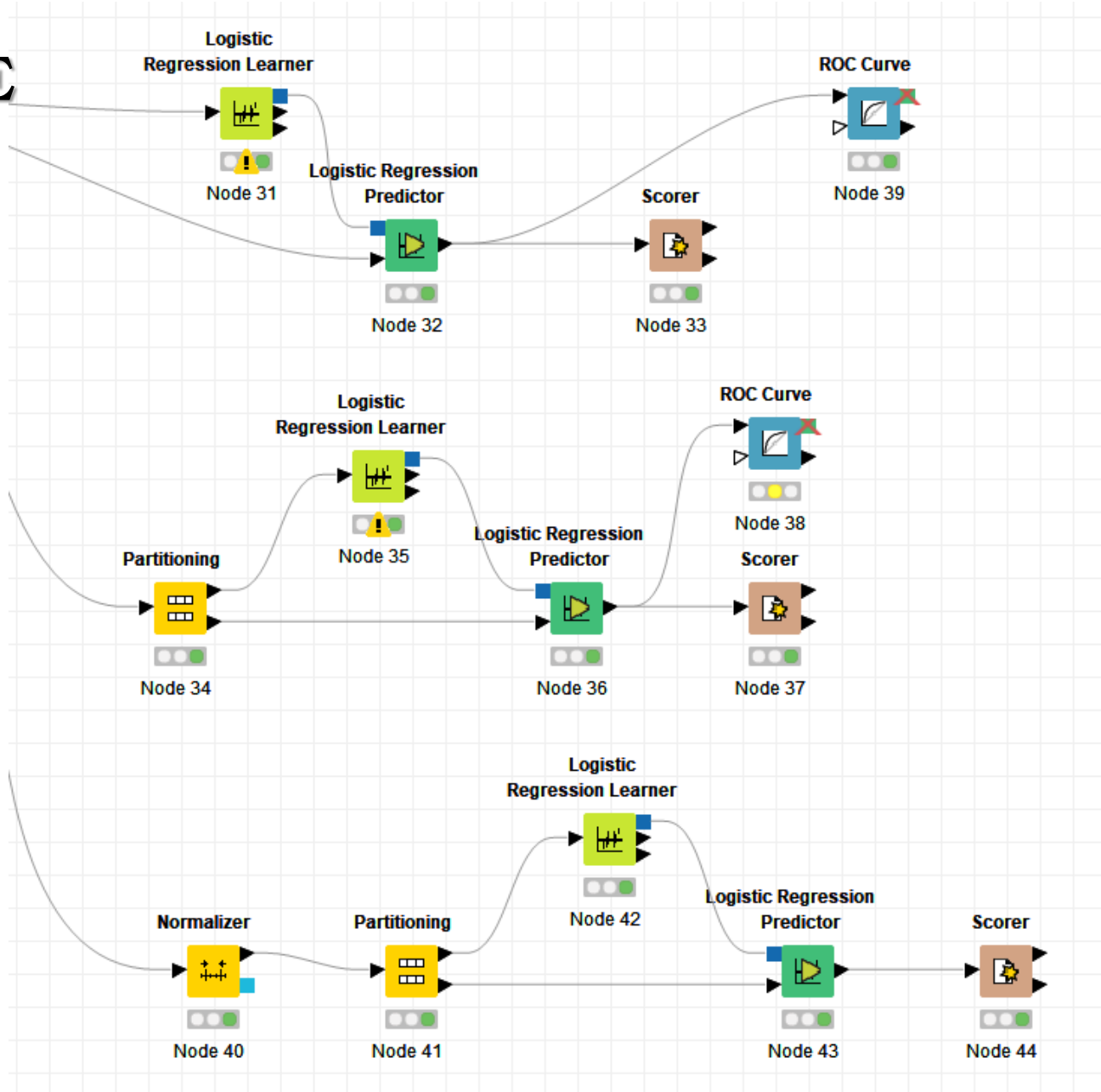


Table: KNIME workflow of Logistic Regression

4. Evaluation & Comparison

► Best Accuracy-wise

- Random Forest: 85.238 %
- Logistic Regression: 85,567 %

► Computation power-wise

- Random Forest: up to 100% CPU utilization in 20+ models
- Logistic Regression: around 20% CPU

► Memory-wise (space complexity)

- Random Forest: up to 10+ GB
- Logistic Regression: around 1+ GB

► Computational time-wise (time complexity)

- Random Forest: generally less than 5 mins
- Logistic Regression: up 25+ minutes for larger number of epochs

4. Discussion & Conclusion

- *Corrections in datasets*
  - Column extraction
  - Filtering
  - Imbalance data

- *Impact of normalization*
  - Random Forest: no effect
  - Logistic Regression: effected

- *Overfitting*
  - Regularization in Random Forest
    e.g. tree depth
  - Regularization in Logistic Regression
    e.g. Laplace

- *Knowledge Discovery in Database* (KDD)

- *Frustration parts*
  - Exhaustive search of best parameter
  - Strange and unexpected accuracies
  - Time effort while trying various values

References

1. Sadi Evren Seker, Lecture Notes, Introduction to Artificial Intelligence and Introduction to Data Science course. (2019 Fall and 2020 Spring)

2. Rain Prediction in Australia Dataset Available at: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.

3. KNIME Analytic Platform. Available at: https://www.knime.com

4. Anaconda Python3 Distribution. Available at: https://www.anaconda.com

Thank You