Data Mining Final Project

By Camilla Sisemore

Attempting to predict the Occupancy of a Room by using Sensor Data

One February, likely in 2015 (based on the publication date of their paper), Luis M. Candanedo and Véronique Feldheim, both of Université de Mons in Mons, Belgium performed an experiment to see if they could determine if a room was occupied based on readings from sensors that measured temperature, humidity, carbon dioxide, and light. The readings were taken three to four times each minute and averaged to give one result per minute per feature. A picture was taken once per minute so that occupancy could be determined by the researchers. The dataset also includes humidity ratio. I will be attempting to determine the occupancy based on combinations of temperature, humidity, carbon dioxide, light, and the humidity ratio. I suspect that CO_2 will definitely be one of the indicators.

[1] (Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models, Energy and Buildings Volume 112, 2016)

The Data

The Train set initially consisted of 8,143 rows and 7 columns. The Train set needed to be cleaned. Every time I thought I had it clean, I was wrong. After cleaning, the train set had 8,140 rows. I used Python with the Pandas library in Jupyter Notebook.

Commands used (some more than once):

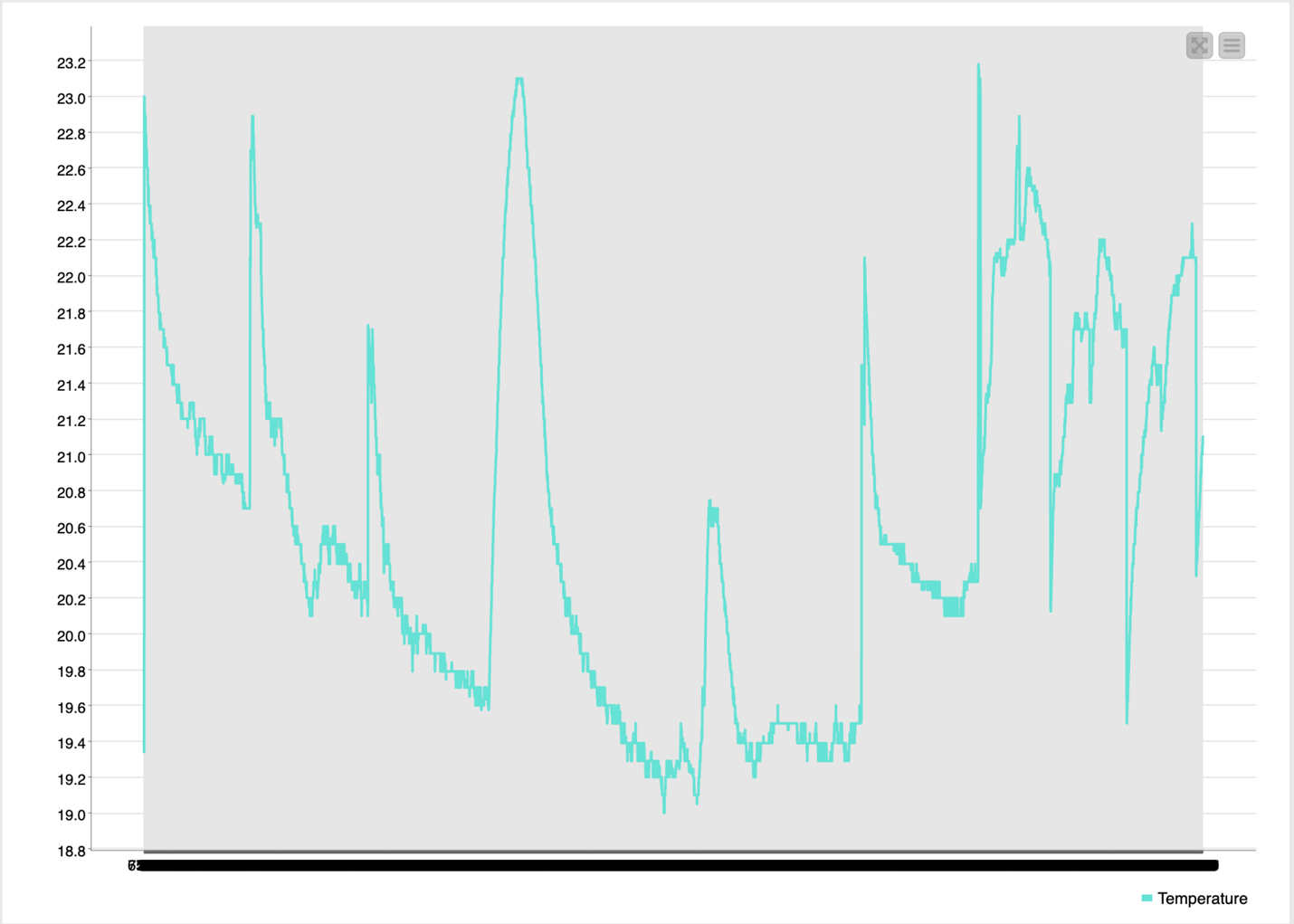
```
df.shape
df.isna().any()
df.dropna(axis = 0, how = 'any', subset=['Occupied'], inplace=True)
df['Occupied'].value_counts()
df['Occupied'] = df['Occupied'].str.replace(' ', ’’)
df['Occupied'] = df['Occupied'].replace('Y', 'yes', regex=True)
df['Occupied'] = df['Occupied'] = df['Occupied'].replace(['N'], 'no’)
df['Occupied'] = df['Occupied'].replace(['yeses'], 'yes’)
df['Occupied'] = df['Occupied'].replace(['No'], 'no')
```

Training Set Exploration

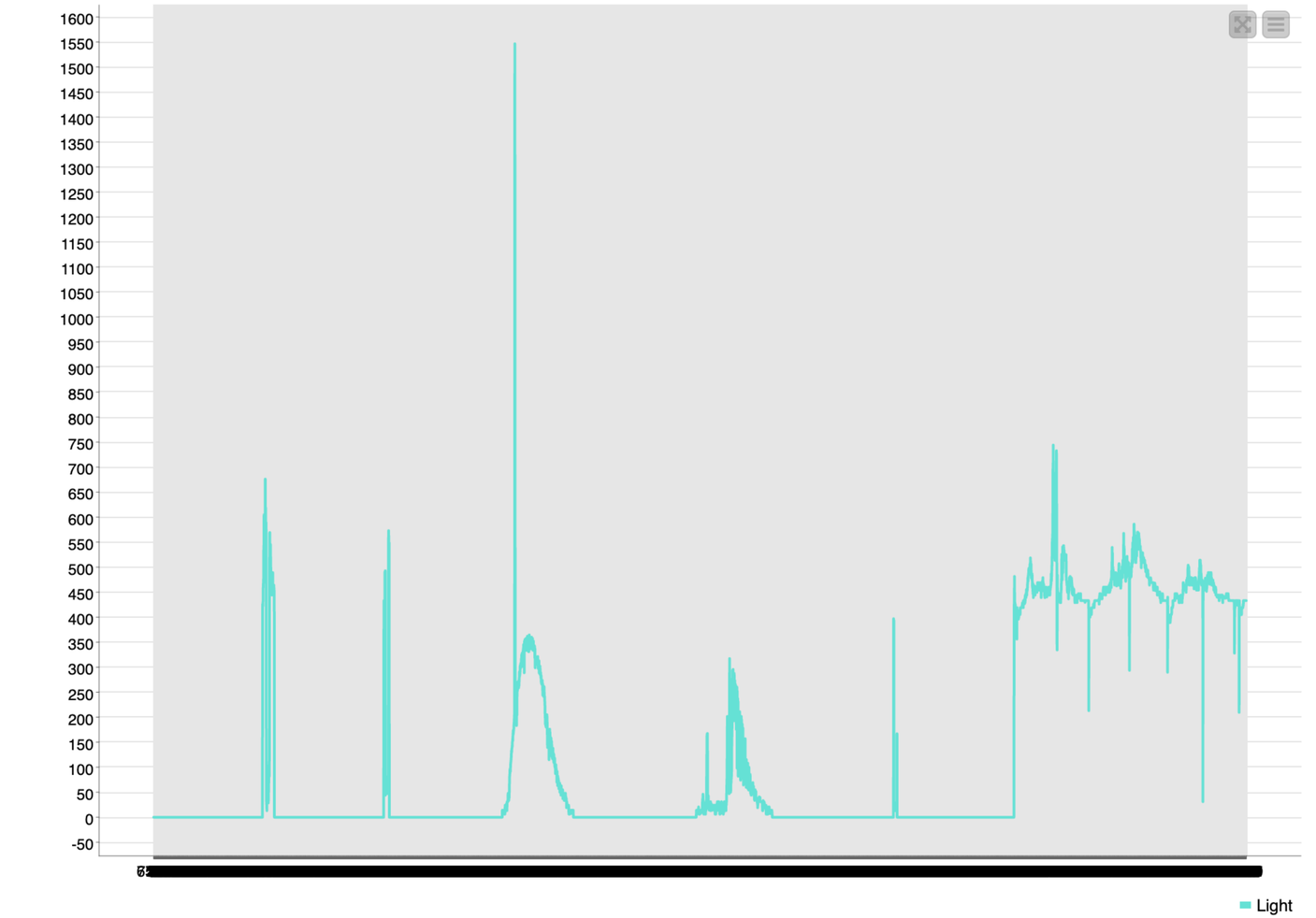
I needed to get the mean and median of the features. I used `df.mean()` and `df.median()` to do this.

Feature	Mean	Median
Temperature	20.618656	20.390000
Humidity	25.731379	26.222500
Light	119.502981	0.000000
CO2	606.533668	453.500000
HumidityRatio	0.003862	0.003801

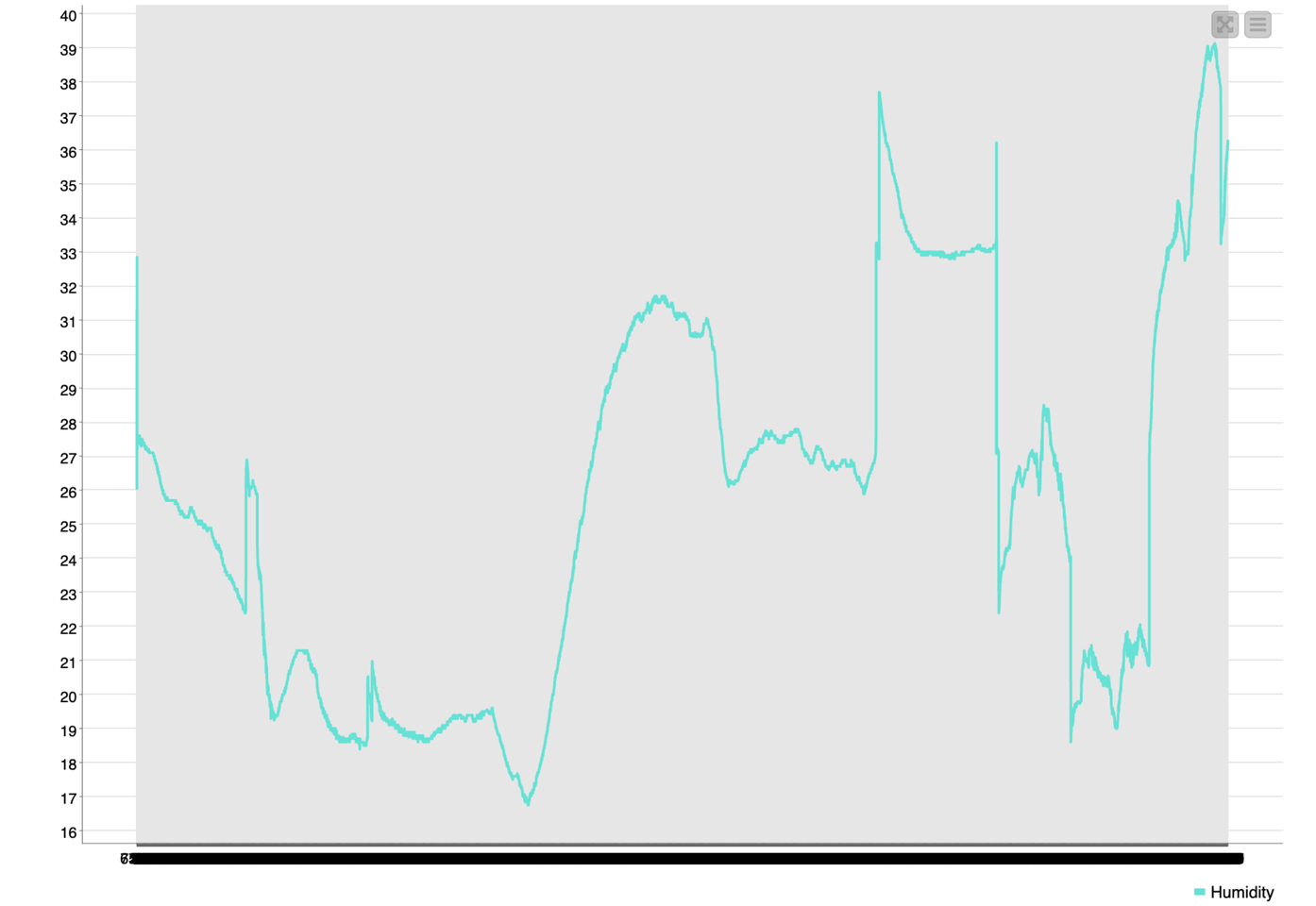
Line Plot - Temperature



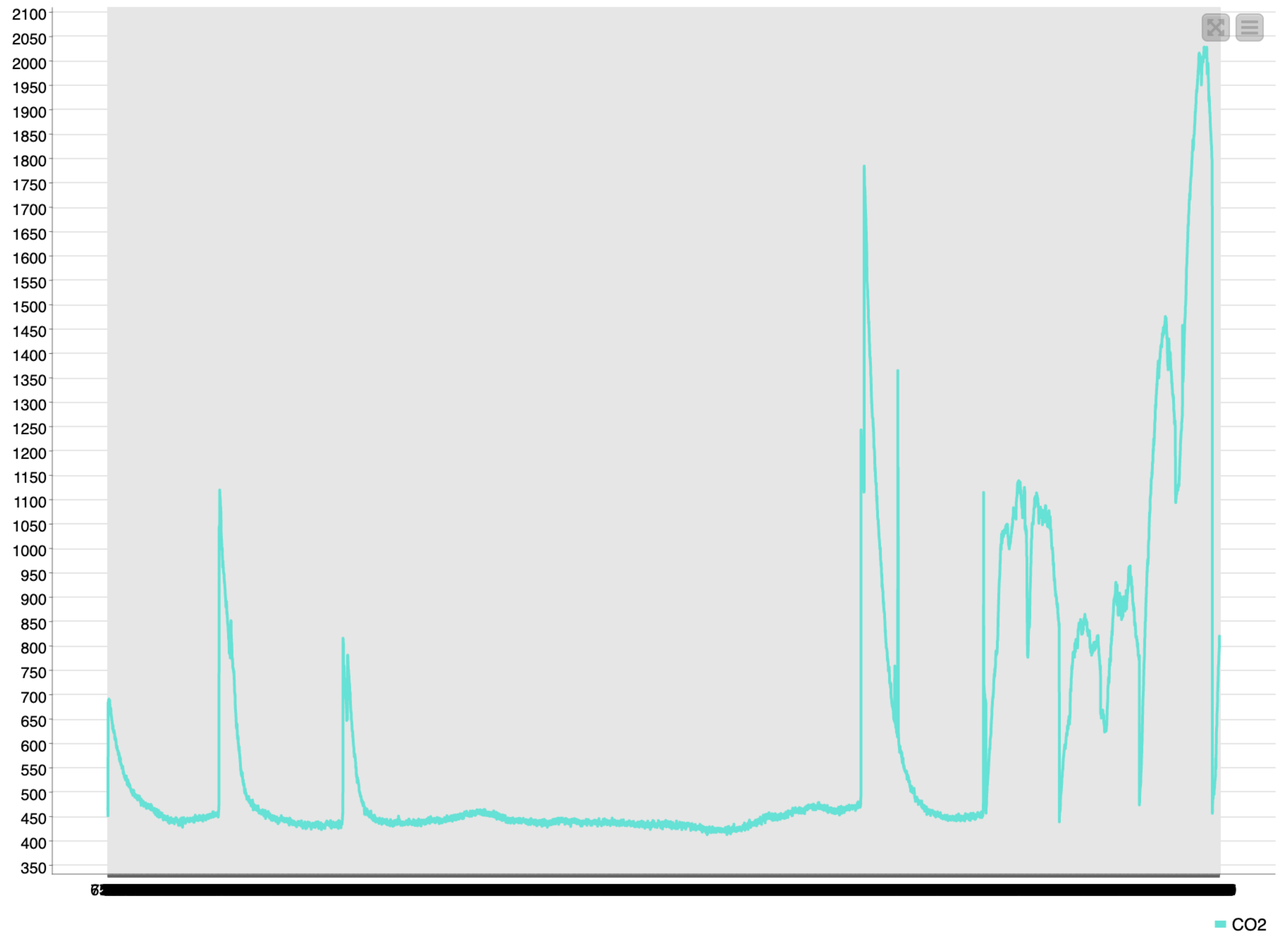
Line Plot - Light



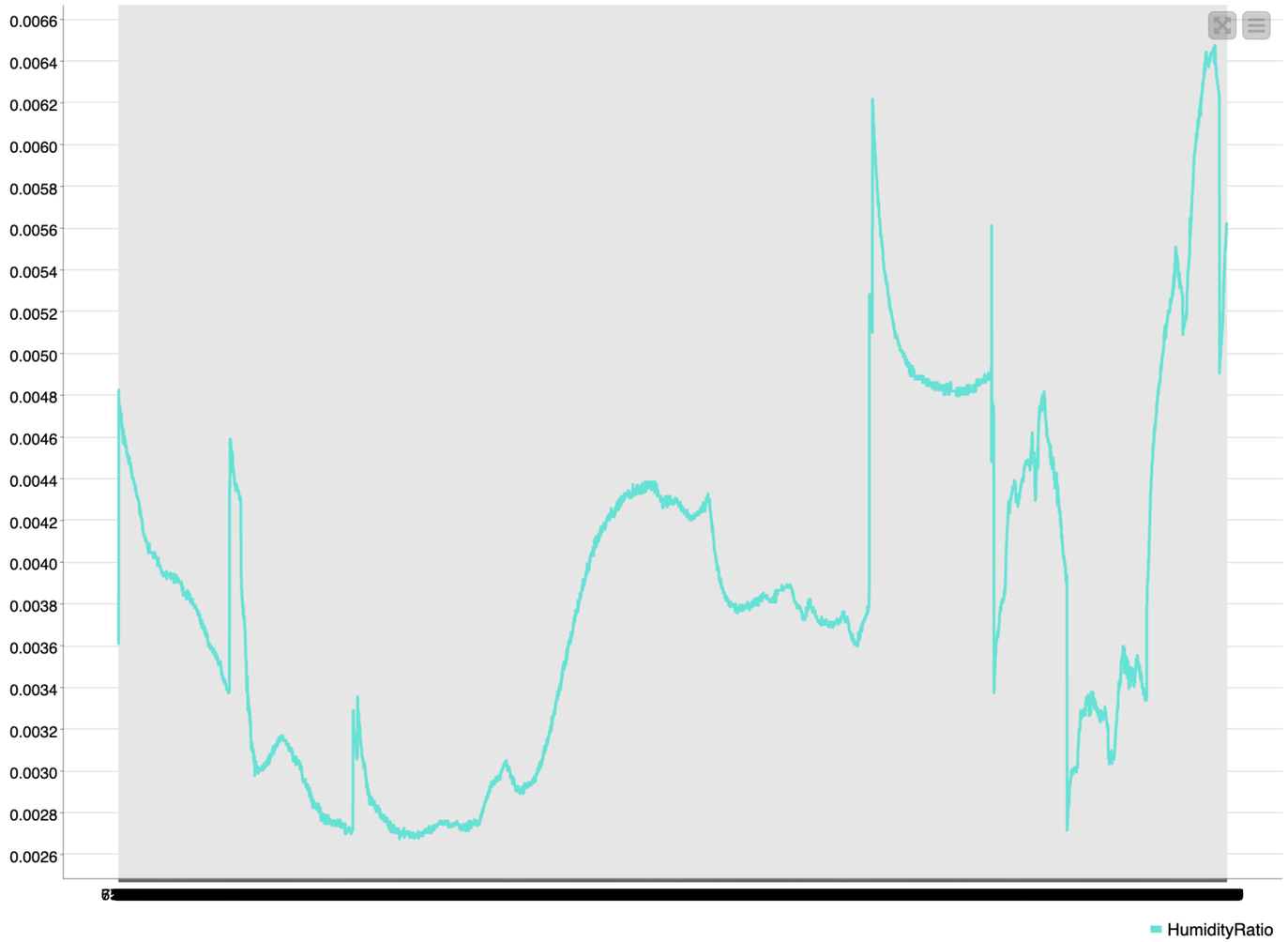
Line Plot - Humidity



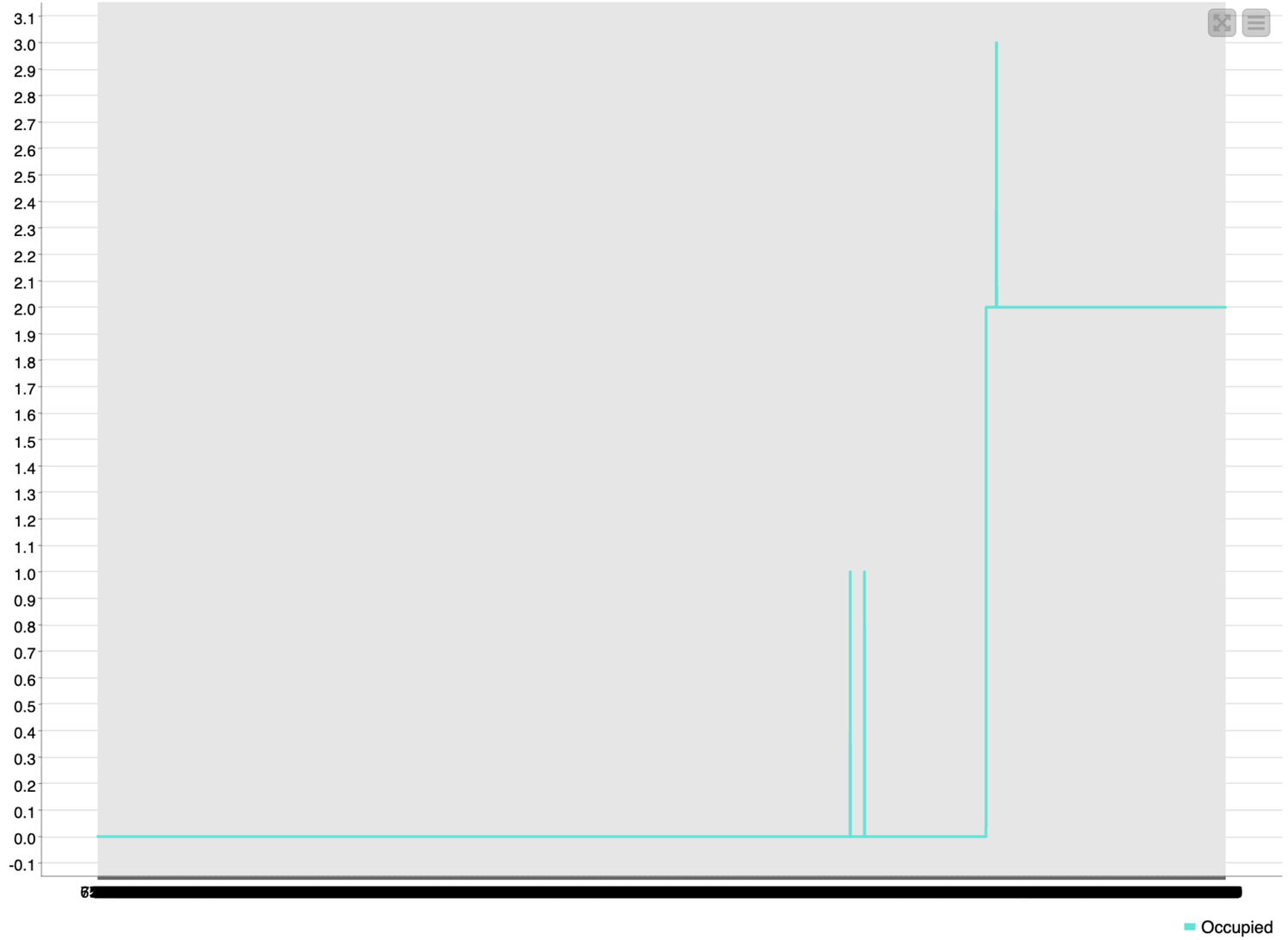
Line Plot – Carbon Dioxide



Line Plot – Humidity Ratio



Line Plot - Occupied

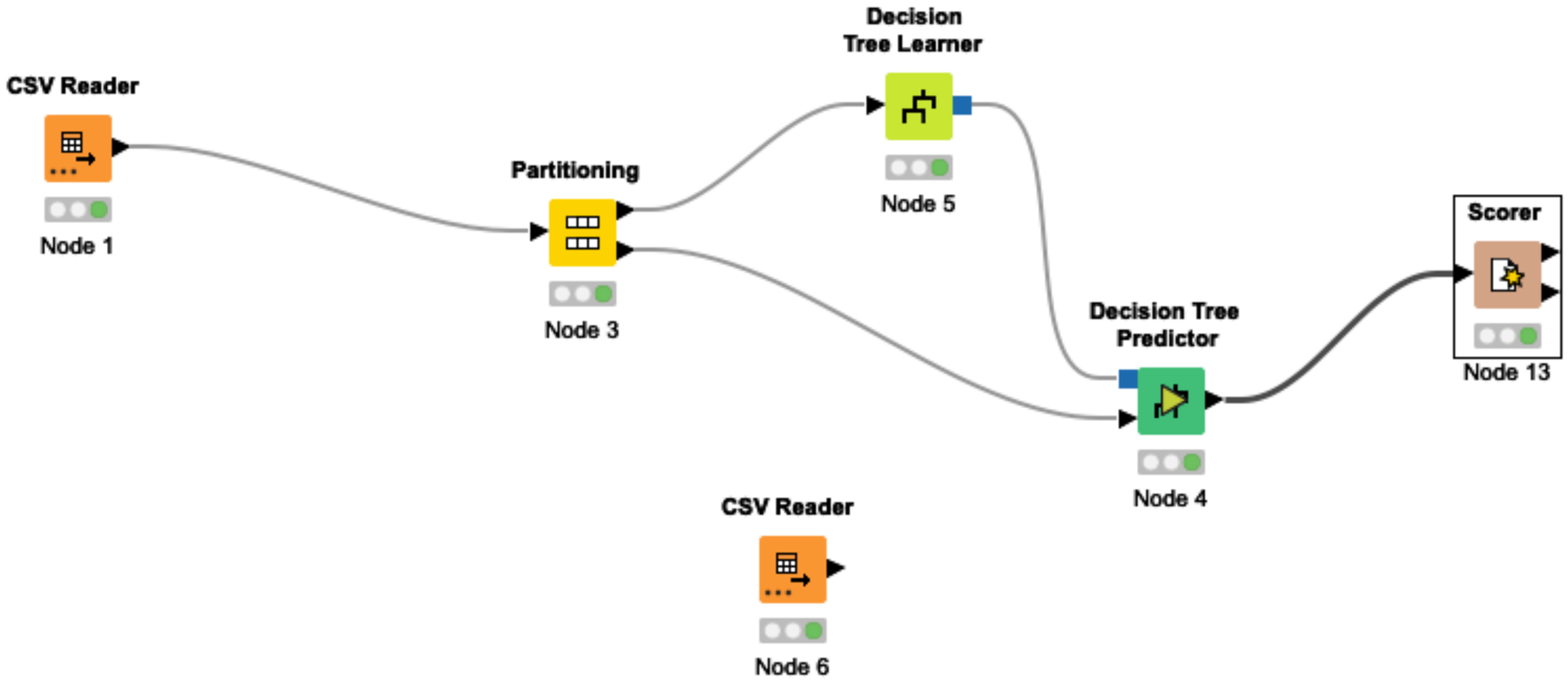


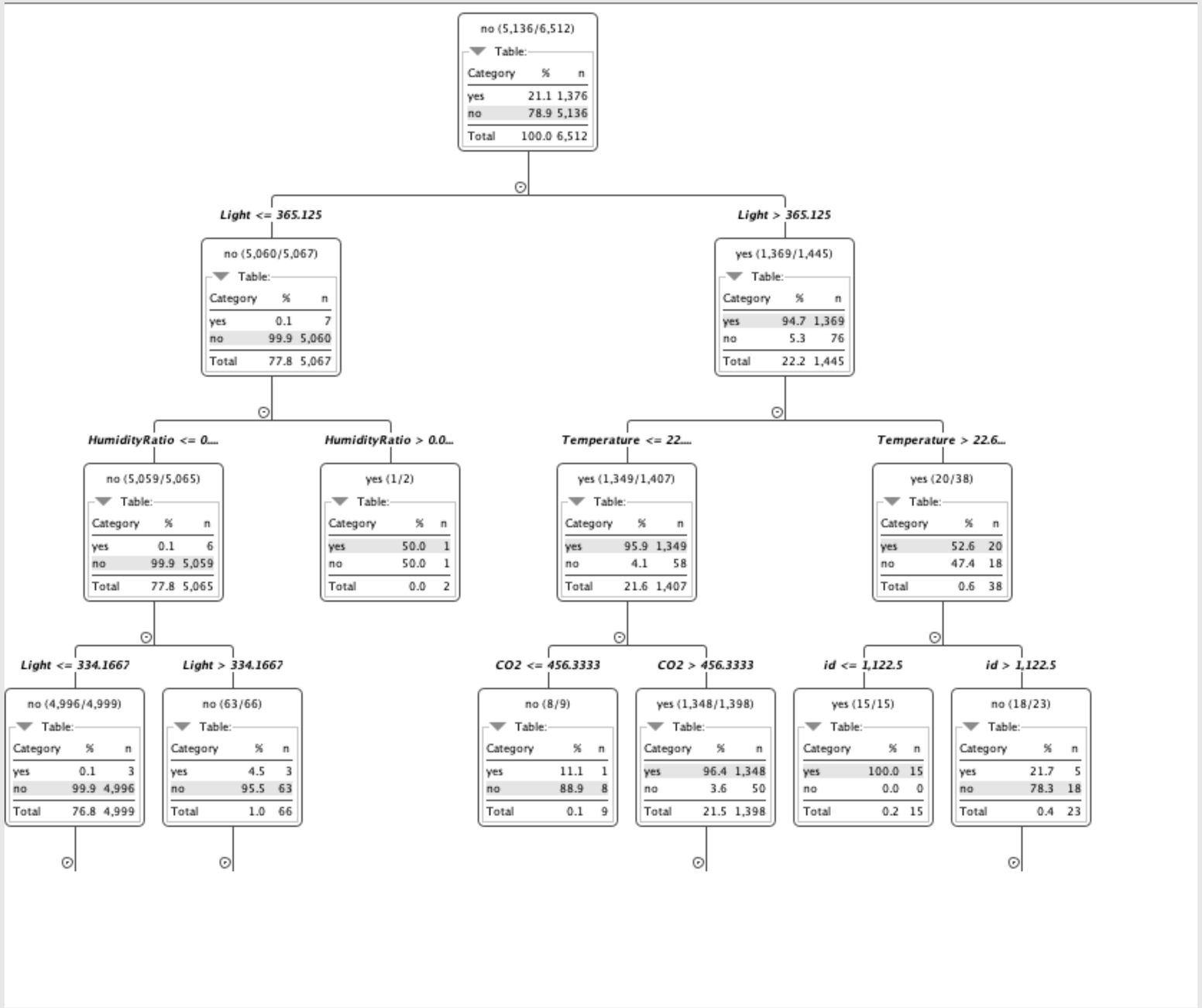
Predicting Occupancy

I will be building models to predict the occupancy of the room. I will be using a train set, which is the dataset that I had to clean, and a test set. The train set is the data that was used to train/teach the model – if only one dataset is provided to work with, the train set typically consumes 70-80% of the dataset.

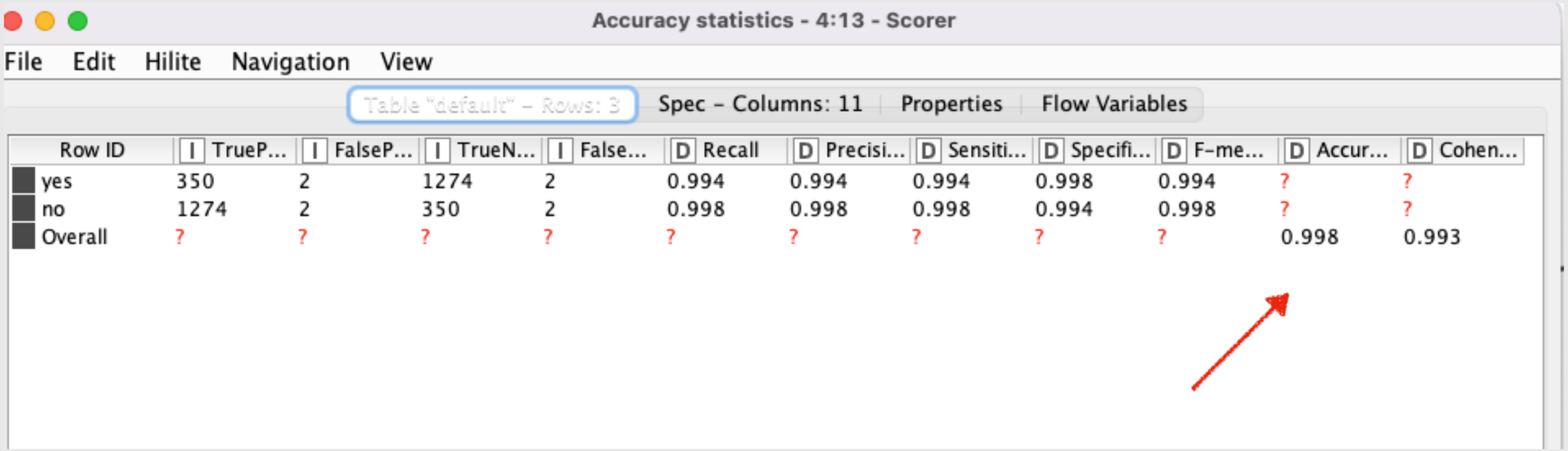
The test set is what is used to test the accuracy of the model created by the train set. If the model is highly accurate, then it can be used again with new data. Due to my inability to figure out the nuances of KNIME and narrowing down the dataset down to specific columns, I manually had to drop columns from the dataset using `df.drop(['Light', 'CO2', 'HumidityRatio'], axis = 1)` as an example.

Decision Tree One: Temperature and Humidity



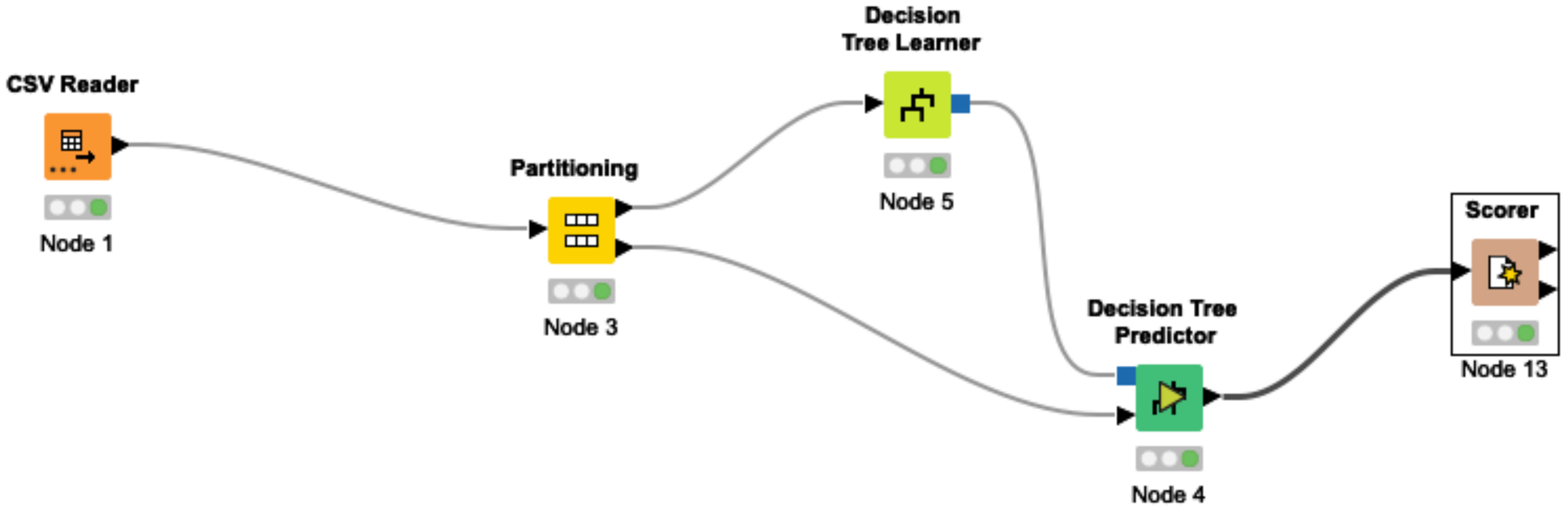

no (5,136/6,512)
Table:
Category % n
yes 21.1 1,376
no 78.9 5,136
Total 100.0 6,512
Light <= 365.125
Light > 365.125
no (5,060/5,067)
Table:
Category % n
yes 0.1 7
no 99.9 5,060
Total 77.8 5,067
yes (1,369/1,445)
Table:
Category % n
yes 94.7 1,369
no 5.3 76
Total 22.2 1,445
HumidityRatio <= 0....
HumidityRatio > 0.0...
Temperature <= 22....
Temperature > 22.6...
no (5,059/5,065)
Table:
Category % n
yes 0.1 6
no 99.9 5,059
Total 77.8 5,065
yes (1/2)
Table:
Category % n
yes 50.0 1
no 50.0 1
Total 0.0 2
yes (1,349/1,407)
Table:
Category % n
yes 95.9 1,349
no 4.1 58
Total 21.6 1,407
yes (20/38)
Table:
Category % n
yes 52.6 20
no 47.4 18
Total 0.6 38
Light <= 334.1667
Light > 334.1667
CO2 <= 456.3333
CO2 > 456.3333
id <= 1,122.5
id > 1,122.5
no (4,996/4,999)
Table:
Category % n
yes 0.1 3
no 99.9 4,996
Total 76.8 4,999
no (63/66)
Table:
Category % n
yes 4.5 3
no 95.5 63
Total 1.0 66
no (8/9)
Table:
Category % n
yes 11.1 1
no 88.9 8
Total 0.1 9
yes (1,348/1,398)
Table:
Category % n
yes 96.4 1,348
no 3.6 50
Total 21.5 1,398
yes (15/15)
Table:
Category % n
yes 100.0 15
no 0.0 0
Total 0.2 15
no (18/23)
Table:
Category % n
yes 21.7 5
no 78.3 18
Total 0.4 23

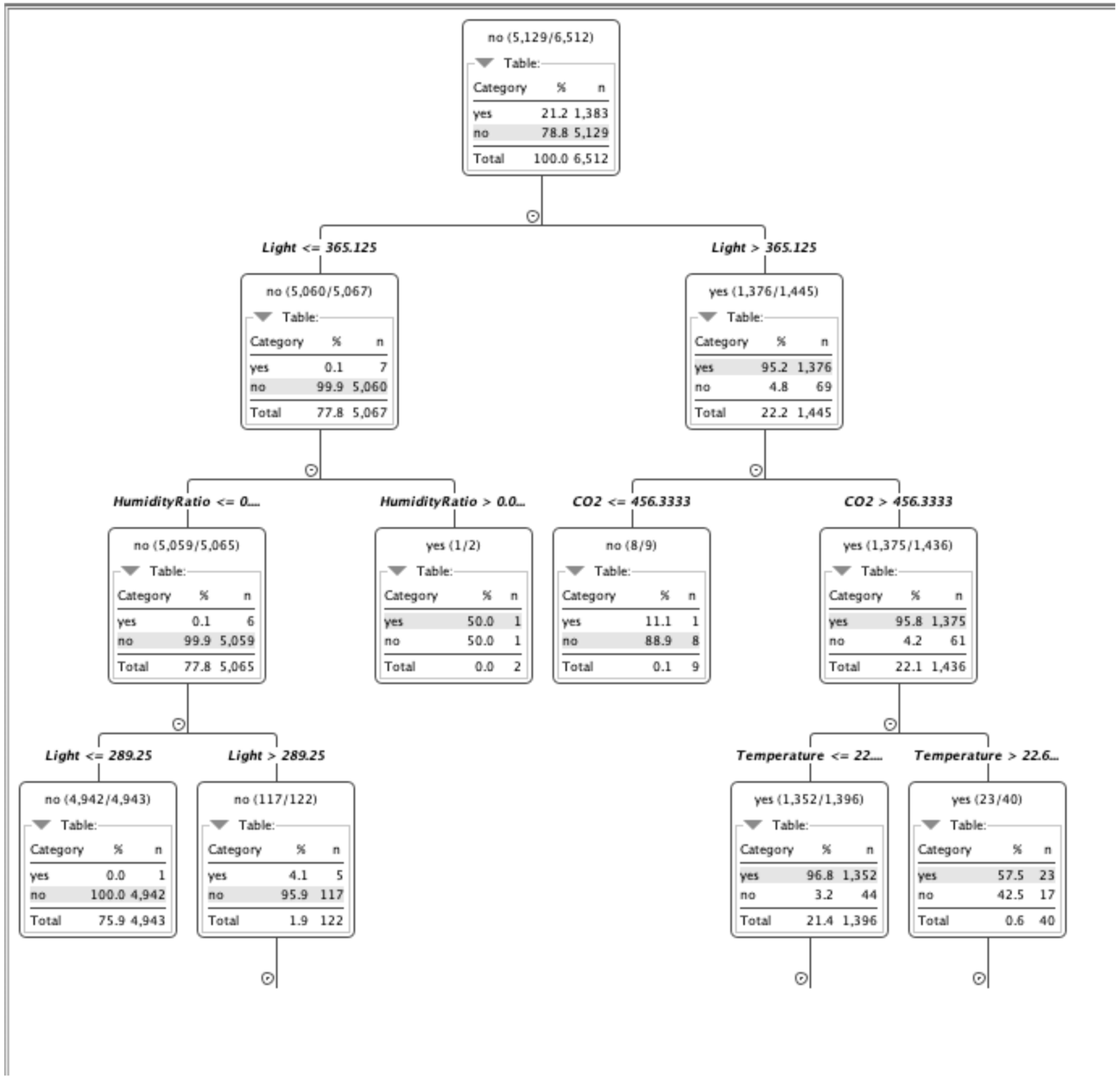
Accuracy



Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifi...	D F-me...	D Accur...	D Cohen...
yes	350	2	1274	2	0.994	0.994	0.994	0.998	0.994	?	?
no	1274	2	350	2	0.998	0.998	0.998	0.994	0.998	?	?
Overall	?	?	?	?	?	?	?	?	?	0.998	0.993

Decision Tree Two: Carbon Dioxide and Humidity Ratio




no (5,129/6,512)
Table:
Category % n
yes 21.2 1,383
no 78.8 5,129
Total 100.0 6,512
Light <= 365.125
Light > 365.125
no (5,060/5,067)
Table:
Category % n
yes 0.1 7
no 99.9 5,060
Total 77.8 5,067
yes (1,376/1,445)
Table:
Category % n
yes 95.2 1,376
no 4.8 69
Total 22.2 1,445
HumidityRatio <= 0....
HumidityRatio > 0.0...
CO2 <= 456.3333
CO2 > 456.3333
no (5,059/5,065)
Table:
Category % n
yes 0.1 6
no 99.9 5,059
Total 77.8 5,065
yes (1/2)
Table:
Category % n
yes 50.0 1
no 50.0 1
Total 0.0 2
no (8/9)
Table:
Category % n
yes 11.1 1
no 88.9 8
Total 0.1 9
yes (1,375/1,436)
Table:
Category % n
yes 95.8 1,375
no 4.2 61
Total 22.1 1,436
Light <= 289.25
Light > 289.25
Temperature <= 22....
Temperature > 22.6...
no (4,942/4,943)
Table:
Category % n
yes 0.0 1
no 100.0 4,942
Total 75.9 4,943
no (117/122)
Table:
Category % n
yes 4.1 5
no 95.9 117
Total 1.9 122
yes (1,352/1,396)
Table:
Category % n
yes 96.8 1,352
no 3.2 44
Total 21.4 1,396
yes (23/40)
Table:
Category % n
yes 57.5 23
no 42.5 17
Total 0.6 40

Accuracy

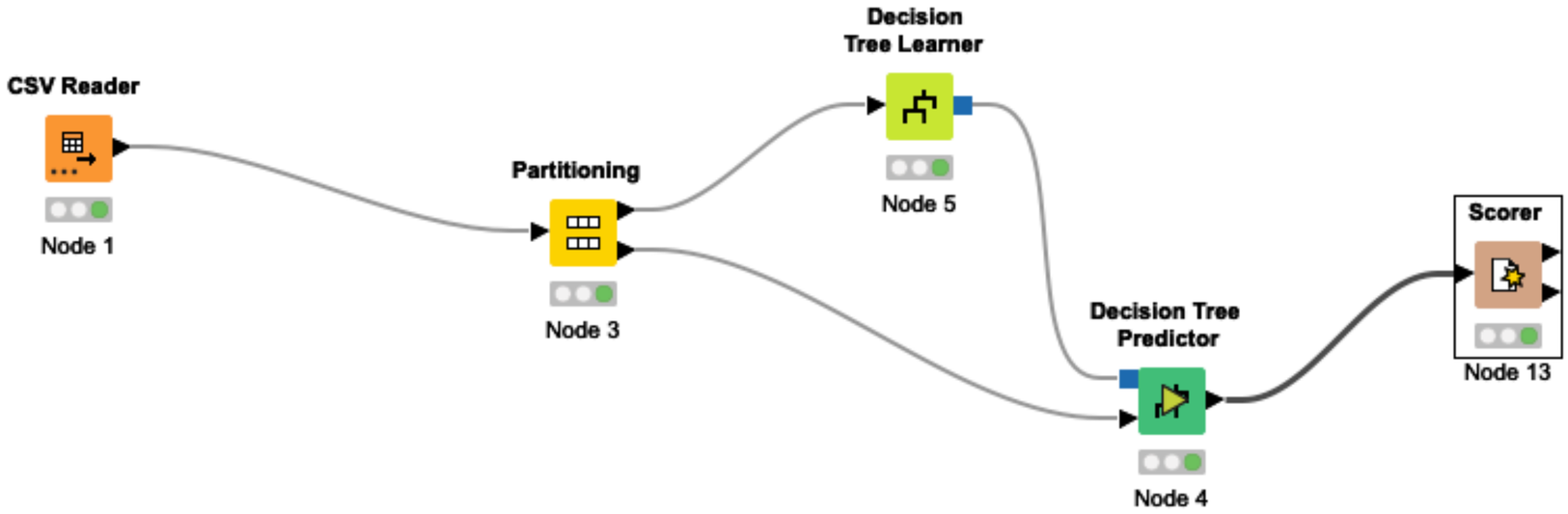
Accuracy statistics - 0:13 - Scorer

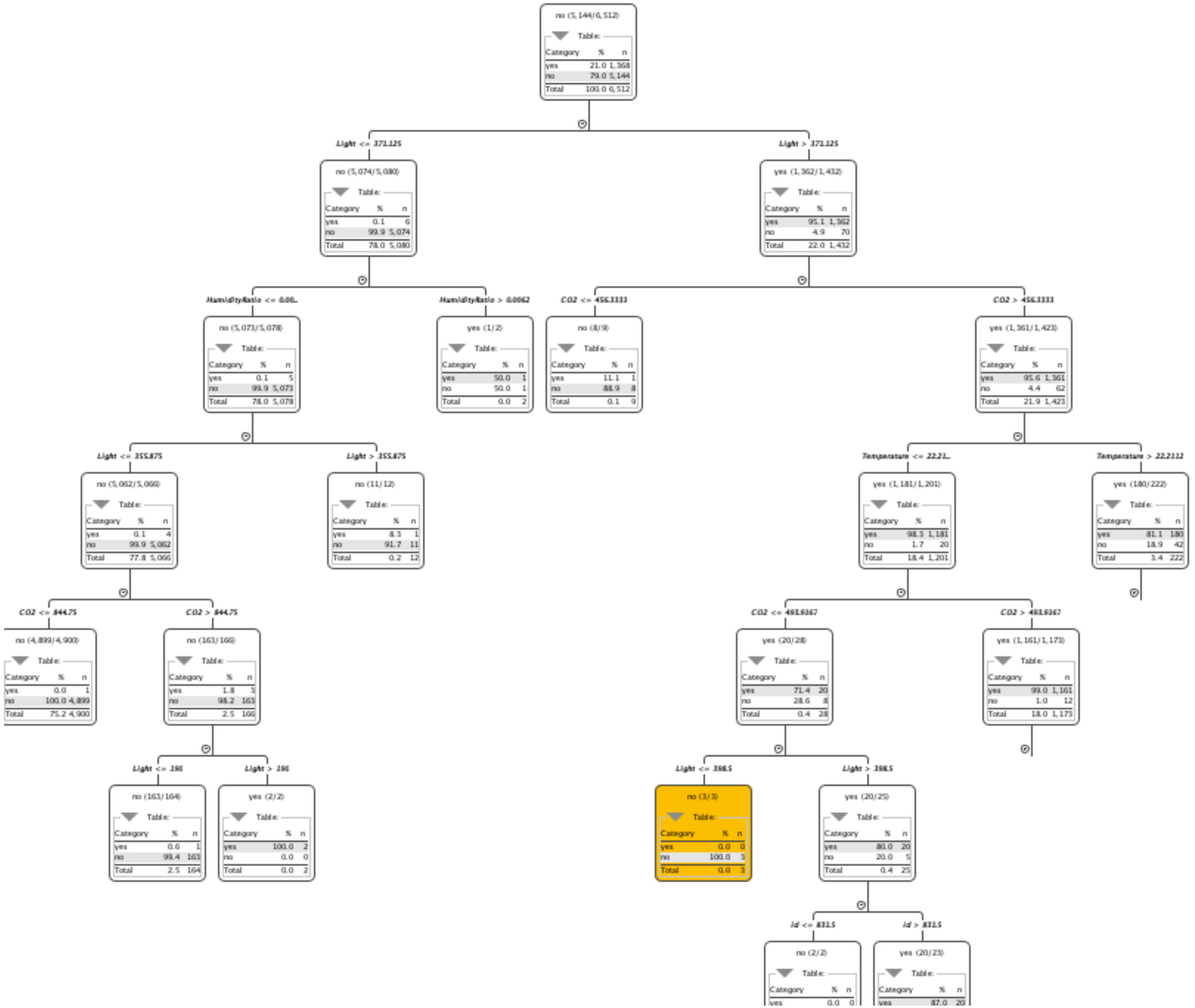
File Edit Hilite Navigation View

Table "default" – Rows: 3 | Spec – Columns: 11 | Properties | Flow Variables

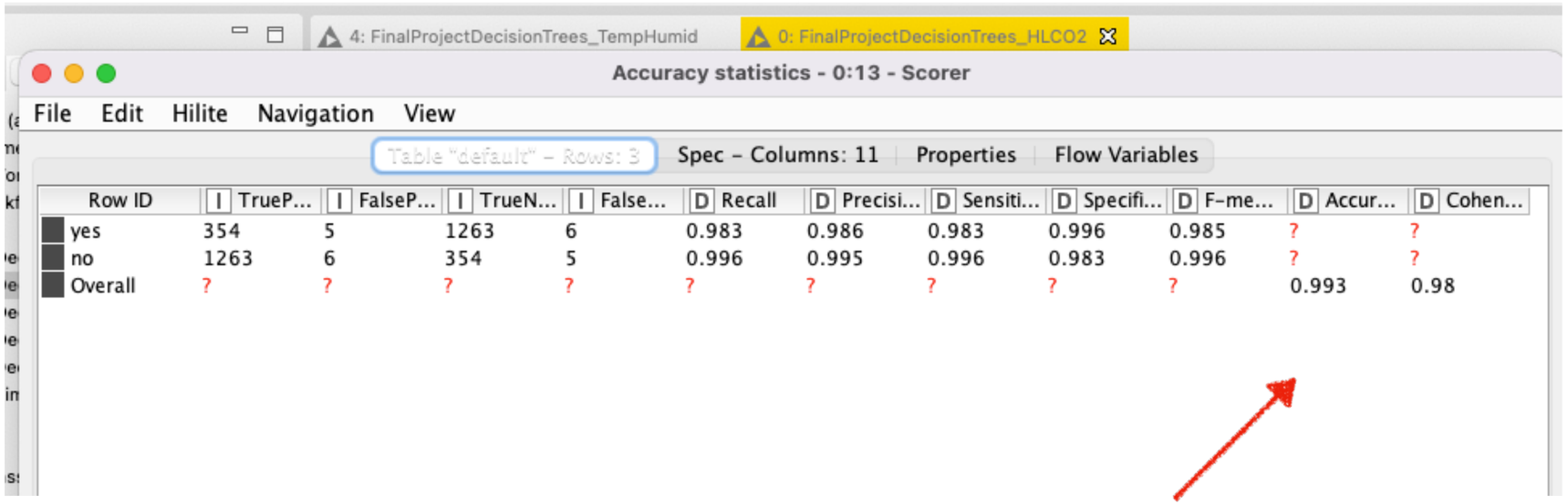
Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifi...	D F-me...	D Accur...	D Cohen...
yes	342	6	1277	3	0.991	0.983	0.991	0.995	0.987	?	?
no	1277	3	342	6	0.995	0.998	0.995	0.991	0.996	?	?
Overall	?	?	?	?	?	?	?	?	?	0.994	0.984

Decision Tree Three: Humidity, Light, and Carbon Dioxide

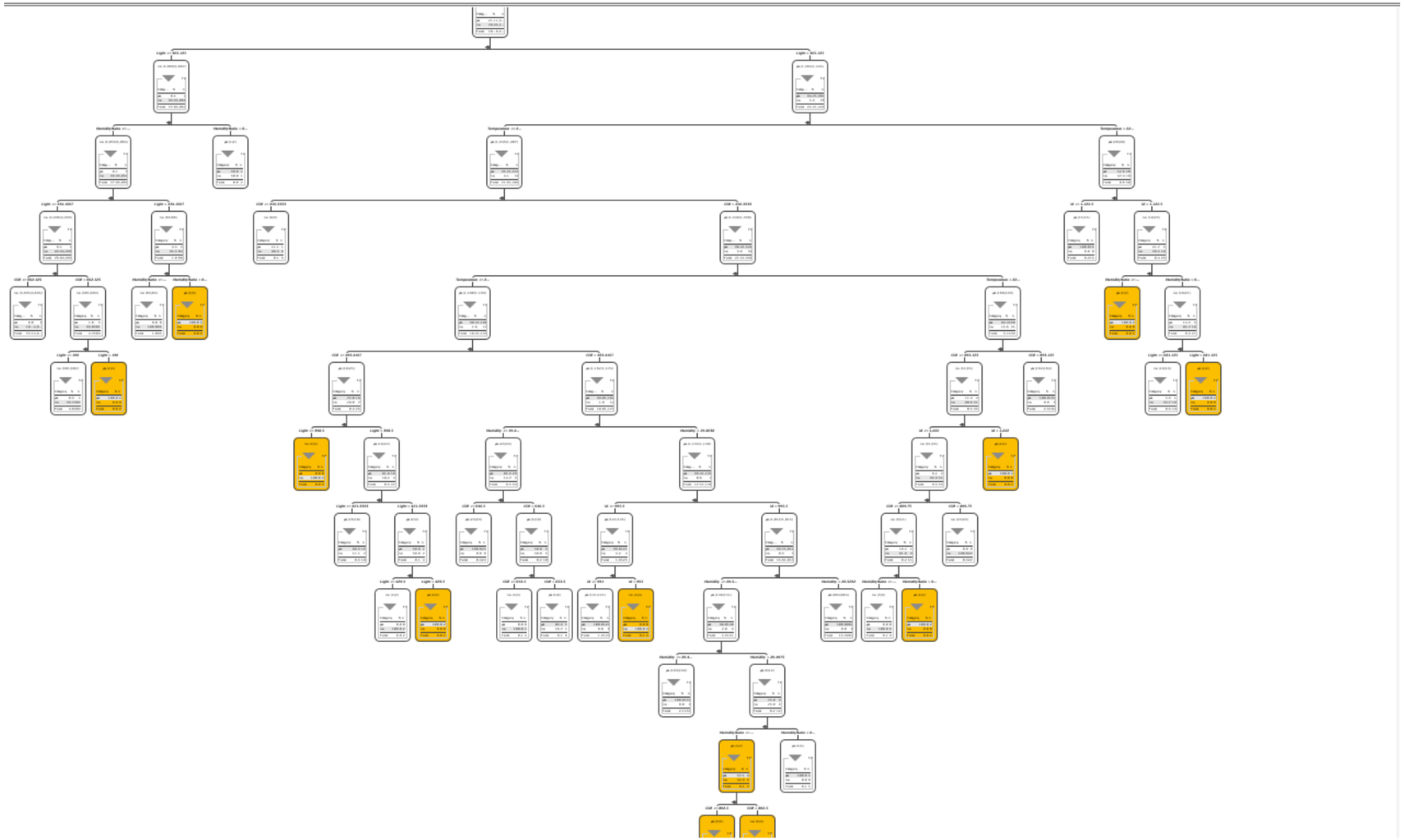



no (5,144/6,512)
Table:
Category % n
yes 21.0 1,368
no 79.0 5,144
Total 100.0 6,512
Light <= 371.125
Light > 371.125
no (5,074/5,080)
Table:
Category % n
yes 0.1 6
no 99.9 5,074
Total 78.0 5,080
yes (1,362/1,432)
Table:
Category % n
yes 95.1 1,362
no 4.9 70
Total 22.0 1,432
HumidityRatio <= 0.00...
HumidityRatio > 0.0062
CO2 <= 456.3333
CO2 > 456.3333
no (5,073/5,078)
Table:
Category % n
yes 0.1 5
no 99.9 5,073
Total 78.0 5,078
yes (1/2)
Table:
Category % n
yes 50.0 1
no 50.0 1
Total 0.0 2
no (8/9)
Table:
Category % n
yes 11.1 1
no 88.9 8
Total 0.1 9
yes (1,361/1,423)
Table:
Category % n
yes 95.6 1,361
no 4.4 62
Total 21.9 1,423
Light <= 355.875
Light > 355.875
Temperature <= 22.21...
Temperature > 22.2112
no (5,062/5,066)
Table:
Category % n
yes 0.1 4
no 99.9 5,062
Total 77.8 5,066
no (11/12)
Table:
Category % n
yes 8.3 1
no 91.7 11
Total 0.2 12
yes (1,181/1,201)
Table:
Category % n
yes 98.3 1,181
no 1.7 20
Total 18.4 1,201
yes (180/222)
Table:
Category % n
yes 81.1 180
no 18.9 42
Total 3.4 222
CO2 <= 844.75
CO2 > 844.75
CO2 <= 493.9167
CO2 > 493.9167
no (4,899/4,900)
Table:
Category % n
yes 0.0 1
no 100.0 4,899
Total 75.2 4,900
no (163/166)
Table:
Category % n
yes 1.8 3
no 98.2 163
Total 2.5 166
yes (20/28)
Table:
Category % n
yes 71.4 20
no 28.6 8
Total 0.4 28
yes (1,161/1,173)
Table:
Category % n
yes 99.0 1,161
no 1.0 12
Total 18.0 1,173
Light <= 396
Light > 396
Light <= 398.5
Light > 398.5
no (163/164)
Table:
Category % n
yes 0.6 1
no 99.4 163
Total 2.5 164
yes (2/2)
Table:
Category % n
yes 100.0 2
no 0.0 0
Total 0.0 2
no (3/3)
Table:
Category % n
yes 0.0 0
no 100.0 3
Total 0.0 3
yes (20/25)
Table:
Category % n
yes 80.0 20
no 20.0 5
Total 0.4 25
Id <= 831.5
Id > 831.5
no (2/2)
Table:
Category % n
yes 0.0 0
yes (20/23)
Table:
Category % n
yes 87.0 20

Accuracy



Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...	Cohen...
yes	354	5	1263	6	0.983	0.986	0.983	0.996	0.985	?	?
no	1263	6	354	5	0.996	0.995	0.996	0.983	0.996	?	?
Overall	?	?	?	?	?	?	?	?	?	0.993	0.98

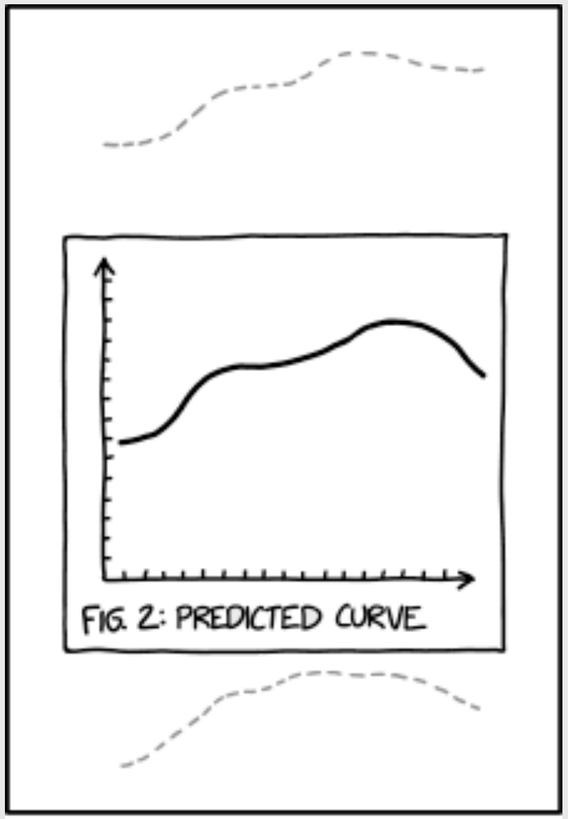


Of these three decision trees, I would have to go with temperature and humidity, as it had the highest accuracy.

Conclusion

I was extremely surprised that the model using temperature and humidity had the highest accuracy, as I had hypothesized to myself that carbon dioxide and humidity ratio would have been higher.

In this project, I cleaned, recleaned, and took the dataset to the dry cleaners. I learned how little I know about KNIME, and that I still have a lot to learn in the Classification area.



SCIENCE TIP: IF YOUR MODEL IS BAD ENOUGH, THE CONFIDENCE INTERVALS WILL FALL OUTSIDE THE PRINTABLE AREA.

https://xkcd.com/2311/