Application

how to use random forest to fill missing data

How to fill the missing data

predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

For example, given that we have many samples that that can be used to train the model and tell if it rains or not

However, for one of the samples there is no wind speed being recorded

How to fill the missing data

predictors

output

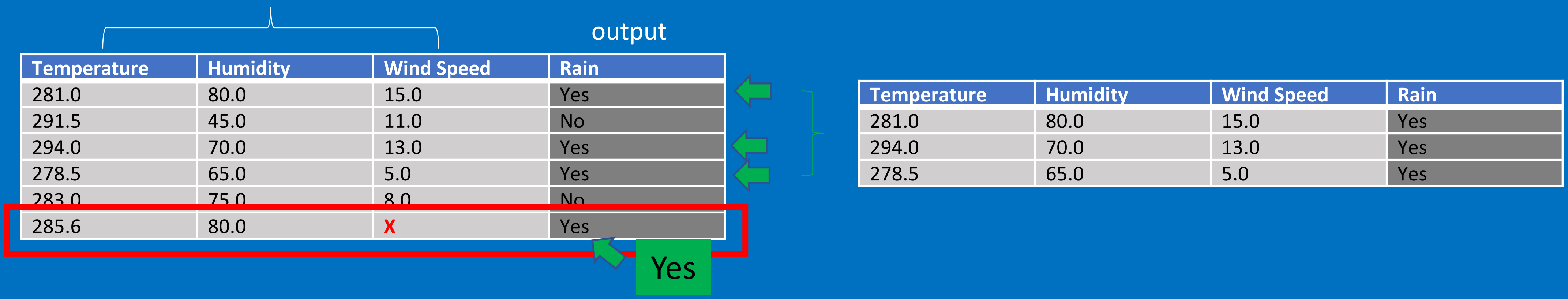
Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

For example, given that we have many samples that that can be used to train the model and tell if it rains or not

However, for one of the samples there is no wind speed being recorded

The general idea of this method is that we first “make a guess” about what the missing data would look like and then gradually tune/refine it to a more optimal value.

How to fill the missing data



predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

In order to fill the missing “wind speed” data **“X”**

Step 1: we locate all the rain value with the same “YES”

How to fill the missing data

predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

Average wind speed = 11.0

In order to fill the missing “wind speed” data “X”

Step 1: we locate all the rain value with the same “YES”

Step 2: For all the sub-selected dataset, the average wind speed is 11.0, so the initial guess of the missing value is 11.0

How to fill the missing data

predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

Average wind speed = 11.0

In order to fill the missing “wind speed” data “X”

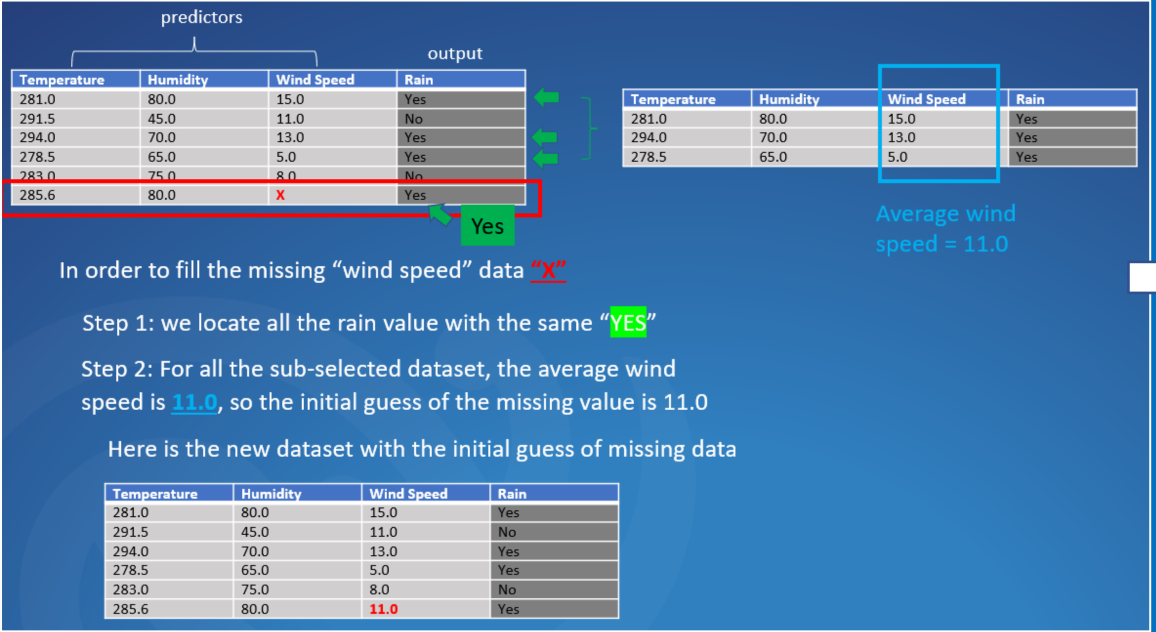
Step 1: we locate all the rain value with the same “YES”

Step 2: For all the sub-selected dataset, the average wind speed is 11.0, so the initial guess of the missing value is 11.0

Here is the new dataset with the initial guess of missing data

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

How to fill the missing data



predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

Average wind speed = 11.0

In order to fill the missing “wind speed” data “X”

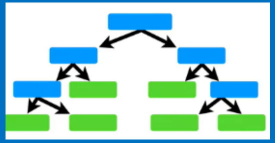
Step 1: we locate all the rain value with the same “YES”

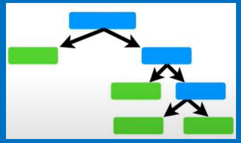
Step 2: For all the sub-selected dataset, the average wind speed is 11.0, so the initial guess of the missing value is 11.0

Here is the new dataset with the initial guess of missing data

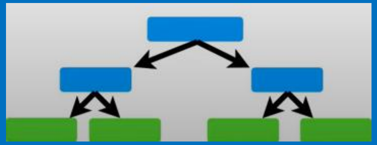
Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Step 3: So using the new dataset, we are able to create a bunch of random forest trees

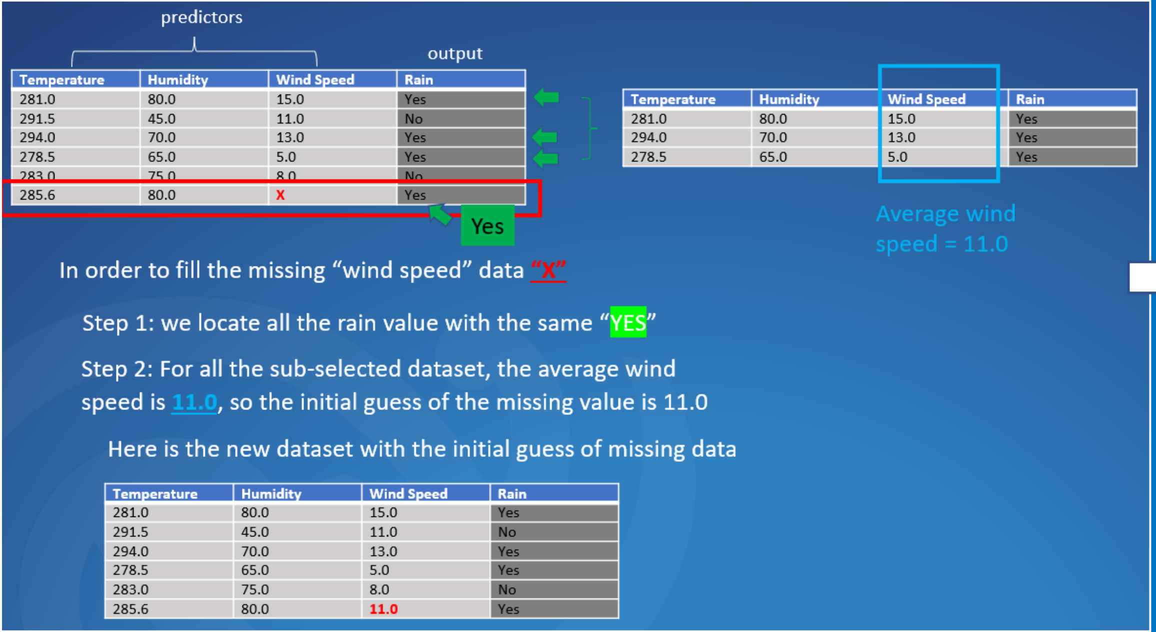




......



How to fill the missing data



predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

Average wind speed = 11.0

In order to fill the missing “wind speed” data “X”

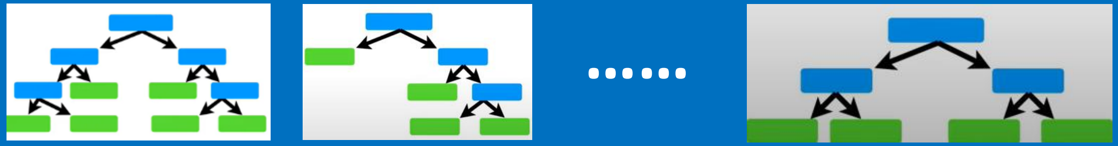
Step 1: we locate all the rain value with the same “YES”

Step 2: For all the sub-selected dataset, the average wind speed is 11.0, so the initial guess of the missing value is 11.0

Here is the new dataset with the initial guess of missing data

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Step 3: So using the new dataset, we are able to create a bunch of random forest trees

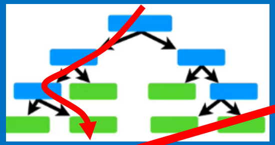


Step 4: We run each dataset down through all the trees individually, and locate similar dataset/samples

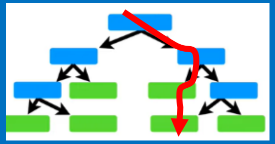
For example, when we use the first tree

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

First dataset



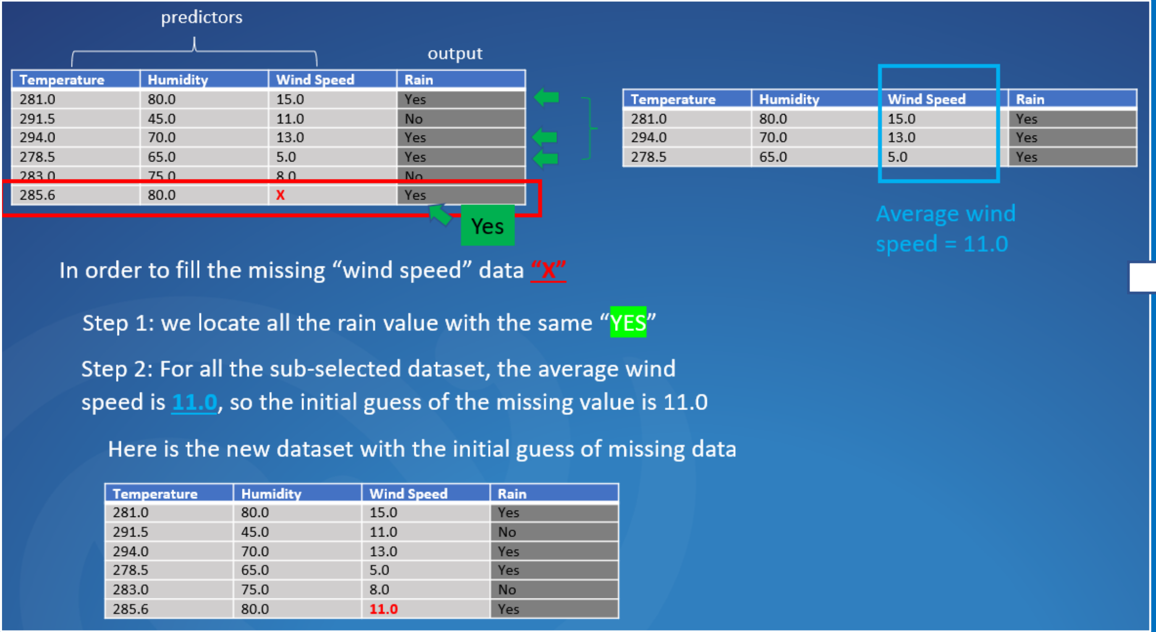
Second dataset



....

As you can see, each dataset/sample will end up at one leaf

How to fill the missing data



predictors

output

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

Average wind speed = 11.0

In order to fill the missing "wind speed" data "X"

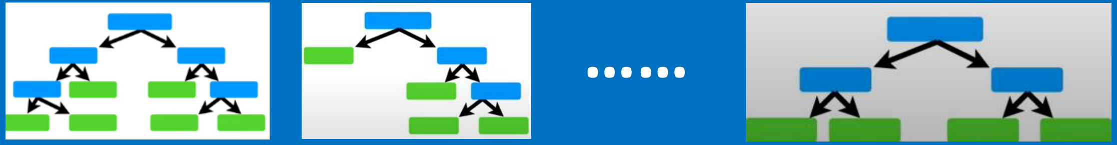
Step 1: we locate all the rain value with the same "YES"

Step 2: For all the sub-selected dataset, the average wind speed is 11.0, so the initial guess of the missing value is 11.0

Here is the new dataset with the initial guess of missing data

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Step 3: So using the new dataset, we are able to create a bunch of random forest trees

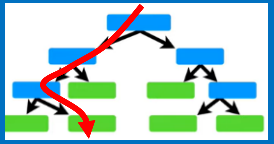


Step 4: We run each dataset down through all the trees individually, and locate similar dataset/samples

For example, when we use the first tree

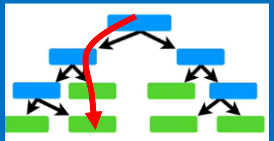
Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

First dataset



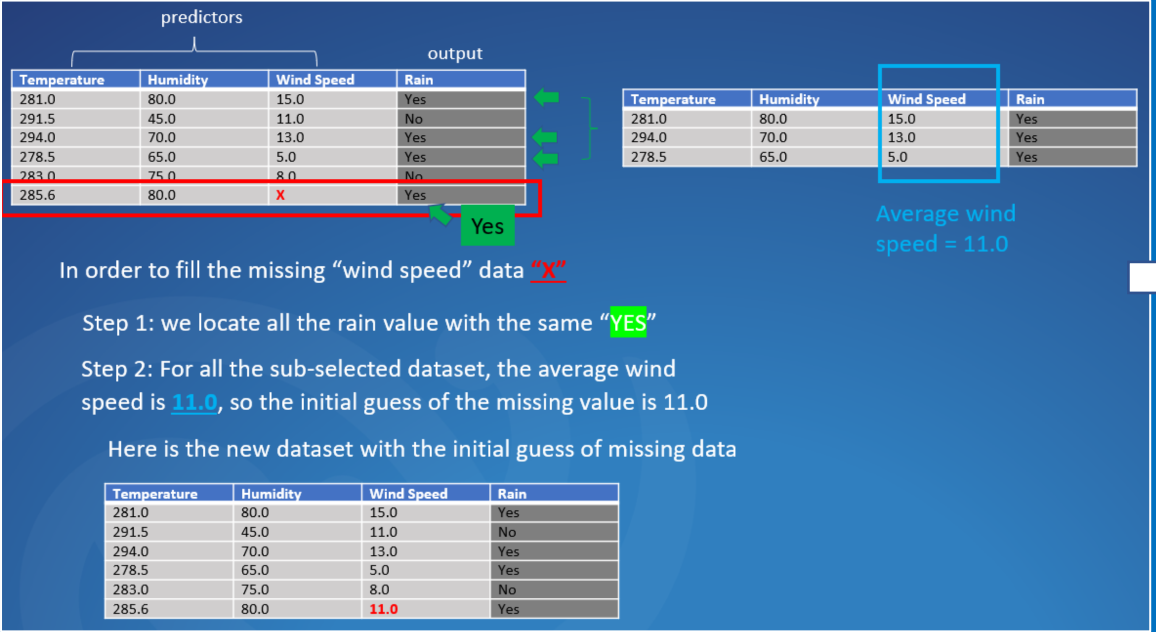
....

4th dataset



The first and the 4th dataset ends up at the same leaf, this means that these two dataset are similar

How to fill the missing data



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	X	Yes

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes

Average wind speed = 11.0

In order to fill the missing “wind speed” data “X”

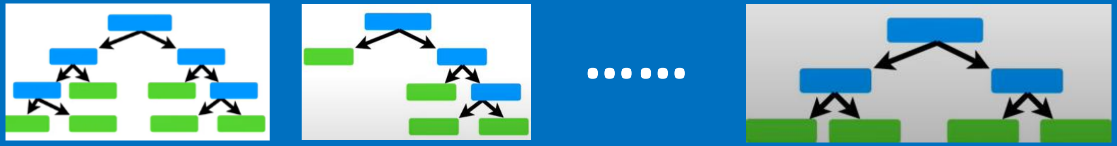
Step 1: we locate all the rain value with the same “YES”

Step 2: For all the sub-selected dataset, the average wind speed is 11.0, so the initial guess of the missing value is 11.0

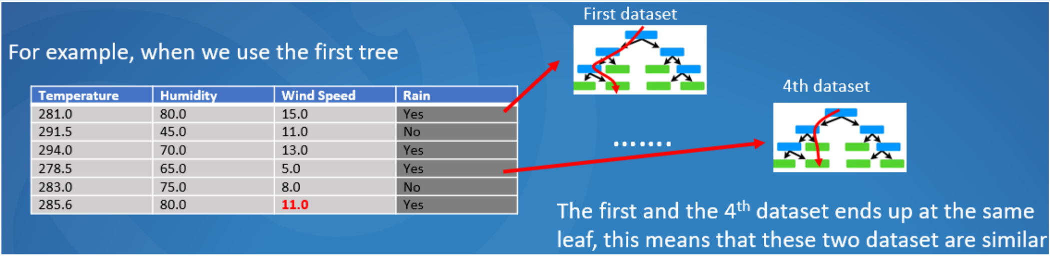
Here is the new dataset with the initial guess of missing data

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Step 3: So using the new dataset, we are able to create a bunch of random forest trees



Step 4: We run each dataset down through all the trees individually, and locate similar dataset/samples



For example, when we use the first tree

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

The first and the 4th dataset ends up at the same leaf, this means that these two dataset are similar

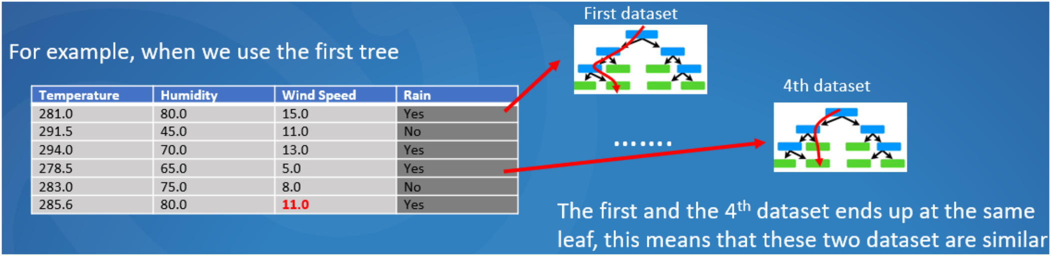
Step 5: we keep track of similar samples using “Proximity matrix”

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

It has both rows/columns equal to the number of samples

How to fill the missing data

Step 4: We run each dataset down through all the trees individually, and locate similar dataset/samples



For example, when we use the first tree

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

The first and the 4th dataset ends up at the same leaf, this means that these two dataset are similar

Step 5: we keep track of similar samples using "Proximity matrix"

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

It has both rows/columns equal to the number of samples

As for this example, when we go through the first tree, the 1st and 4th dataset are similar, so we have

1st sample

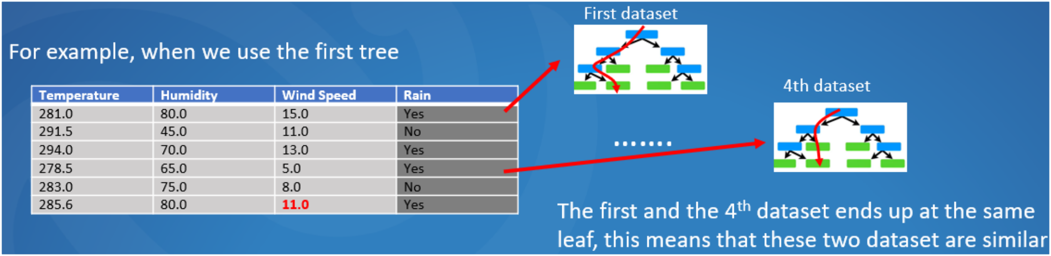
	1	2	3	4	5	6
1						
2						
3						
4	1					
5						
6						

4th sample

We put "1" here to represent the similarity

How to fill the missing data

Step 4: We run each dataset down through all the trees individually, and locate similar dataset/samples



For example, when we use the first tree

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

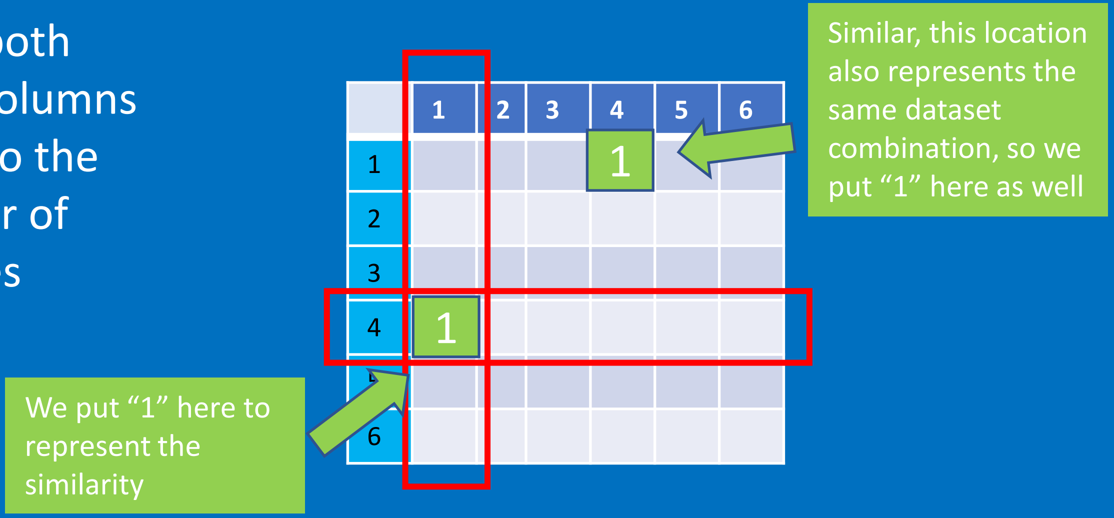
The first and the 4th dataset ends up at the same leaf, this means that these two dataset are similar

Step 5: we keep track of similar samples using "Proximity matrix"

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

It has both rows/columns equal to the number of samples

As for this example, when we go through the first tree, the 1st and 4th dataset are similar, so we have



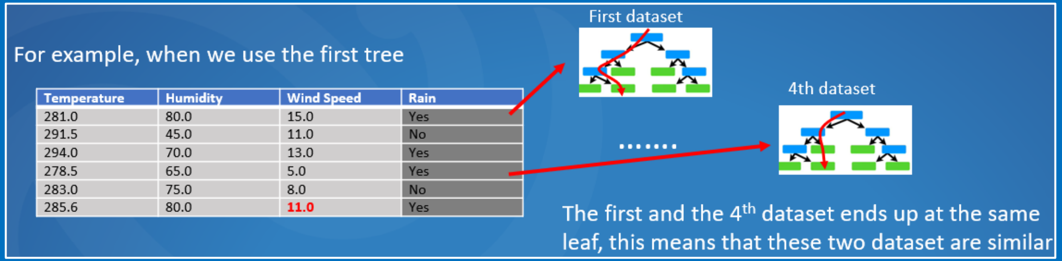
	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Similar, this location also represents the same dataset combination, so we put "1" here as well

We put "1" here to represent the similarity

How to fill the missing data

Step 4: We run each dataset down through all the trees individually, and locate similar dataset/samples



For example, when we use the first tree

Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

The first and the 4th dataset ends up at the same leaf, this means that these two dataset are similar

Step 5: we keep track of similar samples using "Proximity matrix"

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

It has both rows/columns equal to the number of samples

As for this example, when we go through the first tree, the 1st and 4th dataset are similar, so we have

	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Our "Proximity matrix" looks like the left after we go through the first tree

How to fill the missing data

Step 5: we keep track of similar samples using “Proximity matrix”

	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Our “Proximity matrix” looks like the left after we go through the first tree

How to fill the missing data

Step 5: we keep track of similar samples using "Proximity matrix"

	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Our "Proximity matrix" looks like the left after we go through the first tree

	1	2	3	4	5	6
1				2	1	
2						
3				1		
4	2		1			
5						1
6						

Our "Proximity matrix" looks like the left after we go through the 2nd tree (sample 1 and sample 4 ended up the same leaf again …)

How to fill the missing data

Step 5: we keep track of similar samples using “Proximity matrix”

	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Our “Proximity matrix” looks like the left after we go through the first tree

......

	1	2	3	4	5	6
1				8	1	
2			2			
3		2		3	1	
4	8		3			
5			1			1
6						

Ultimately, after gone through all the trees, our “Proximity matrix” looks like the left

How to fill the missing data

Step 5: we keep track of similar samples using “Proximity matrix”

	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Our “Proximity matrix” looks like the left after we go through the first tree

......

	1	2	3	4	5	6
1				0.8		
2			0.2			
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6					0.1	

Ultimately, after gone through all the trees, our “Proximity matrix” looks like the left. And we divide the proximity value by the total number of trees (assuming we have 10 trees).

How to fill the missing data

Step 5: we keep track of similar samples using “Proximity matrix”

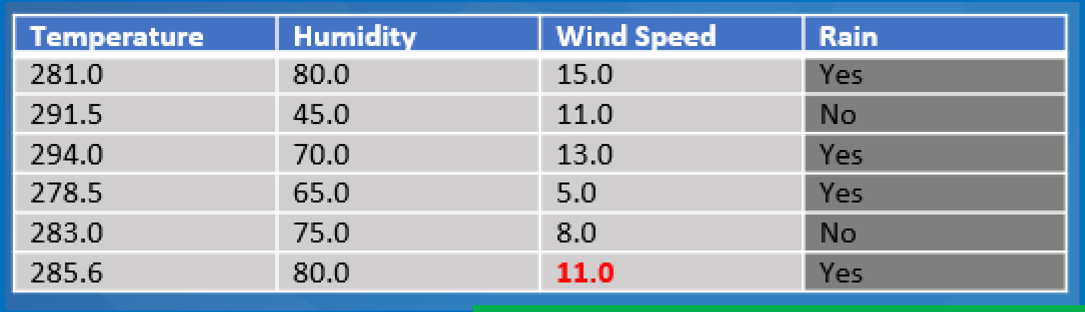
	1	2	3	4	5	6
1				1		
2						
3						
4	1					
5						
6						

Our “Proximity matrix” looks like the left after we go through the first tree

……

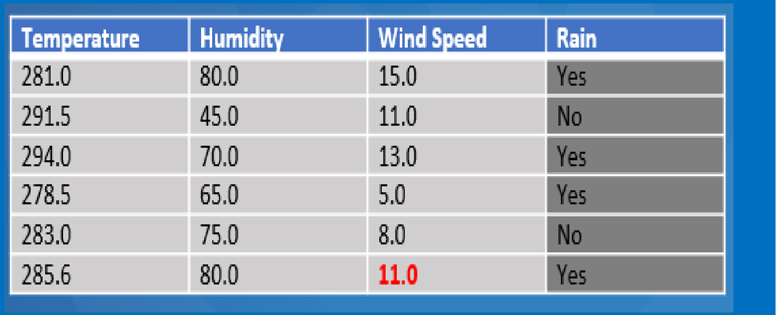
	1	2	3	4	5	6
1				0.8		
2			0.2			
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6					0.1	

Ultimately, after gone through all the trees, our “Proximity matrix” looks like the left. And we divide the proximity value by the total number of trees (assuming we have 10 trees).



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Now we can use the “Proximity matrix” to make a better guess for the **missing value**



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

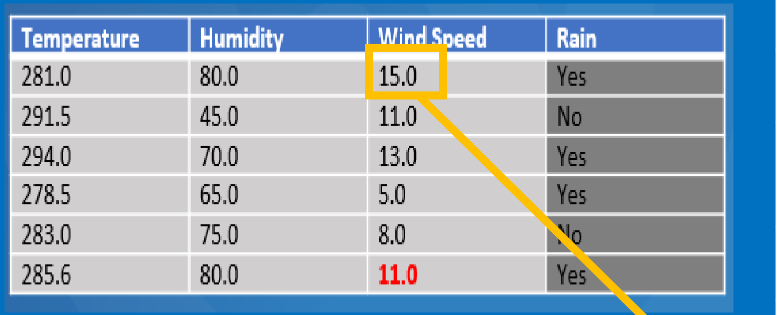
	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

Sample 6

Weight function

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

Sample 6

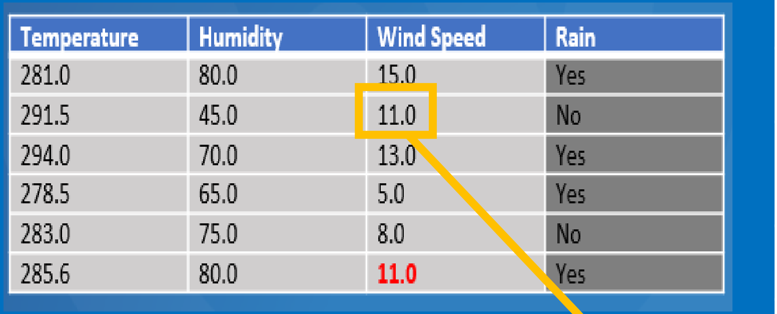
Weight function

0.0 0.3 0.1

The weight for sample 1 $w_1 = 15.0 \times \frac{0.0}{0.3 + 0.1} = 0.0$

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

Sample 6

Weight function

0.3

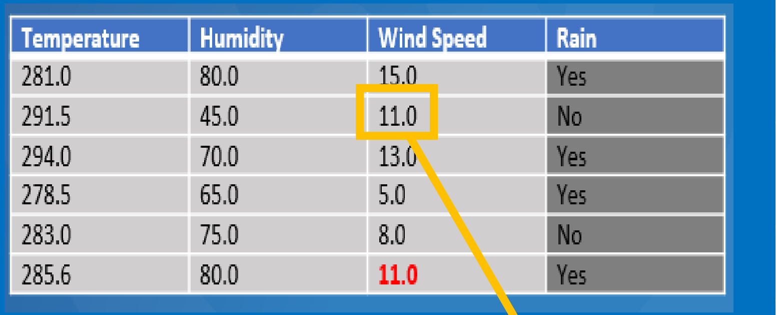
0.1

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)

The weight for sample 1 $w_1 = 15.0 \times \frac{0.0}{0.3 + 0.1} = 0.0$

The weight for sample 2 $w_2 = 11.0 \times \frac{0.3}{0.3 + 0.1} = 8.25$



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

Sample 6

Weight function

0.3

0.1

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)

The weight for sample 1

$$w_1 = 15.0 \times \frac{0.0}{0.3 + 0.1} = 0.0$$

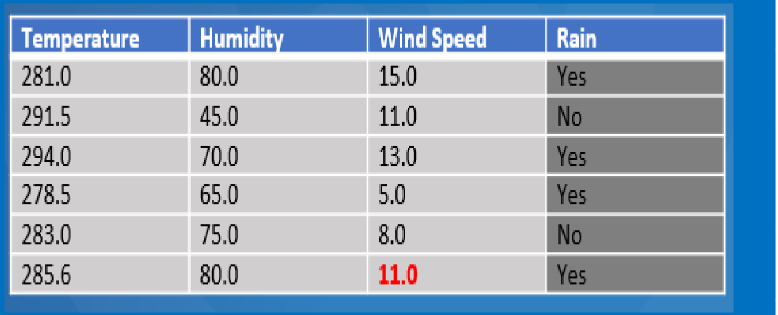
The weight for sample 2

$$w_2 = 11.0 \times \frac{0.3}{0.3 + 0.1} = 8.25$$

……

The weight for sample 5

$$w_5 = 8.0 \times \frac{0.1}{0.3 + 0.1} = 2.0$$



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

Sample 6

0.3 0.1

Weight function

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)

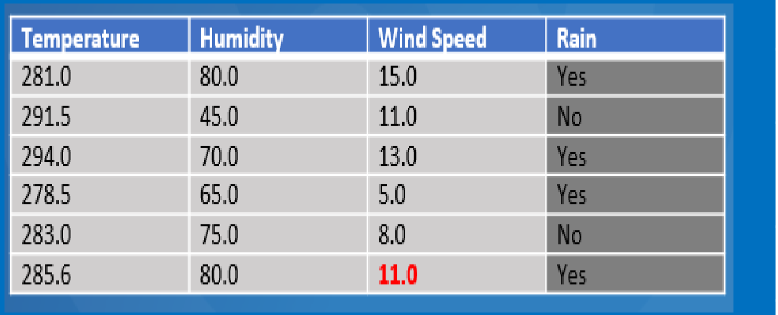
The weight for sample 1 $w_1 = 15.0 \times \frac{0.0}{0.3 + 0.1} = 0.0$

The weight for sample 2 $w_2 = 11.0 \times \frac{0.3}{0.3 + 0.1} = 8.25$

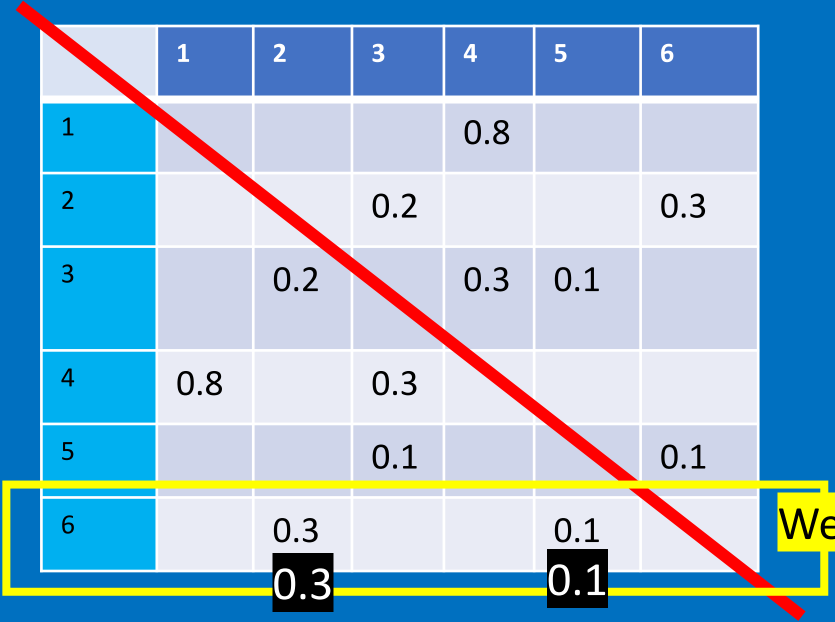
……

The weight for sample 5 $w_5 = 8.0 \times \frac{0.1}{0.3 + 0.1} = 2.0$

The weight for sample 6 $w_6 = 0.0$



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes



	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

0.3 0.1

Sample 6

Weight function

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)

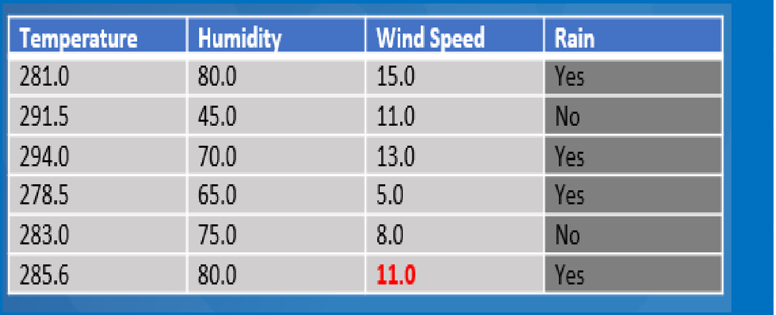
The weight for sample 1 $$w_1 = 15.0 \times \frac{0.0}{0.3 + 0.1} = 0.0$$

The weight for sample 2 $$w_2 = 11.0 \times \frac{0.3}{0.3 + 0.1} = 8.25$$

……

The weight for sample 5 $$w_5 = 8.0 \times \frac{0.1}{0.3 + 0.1} = 2.0$$

The weight for sample 6 $$w_6 = 0.0$$



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

	1	2	3	4	5	6
1				0.8		
2			0.2			0.3
3		0.2		0.3	0.1	
4	0.8		0.3			
5			0.1			0.1
6		0.3			0.1	

Sample 6

0.3 0.1

Weight function

Note that

1. the proximity matrix is a diagonal matrix,
2. the shaded area indicate the weight function for different samples against sample 6 (which has the missing value)

The weight for sample 1 $$w_1 = 15.0 \times \frac{0.0}{0.3+0.1} = 0.0$$

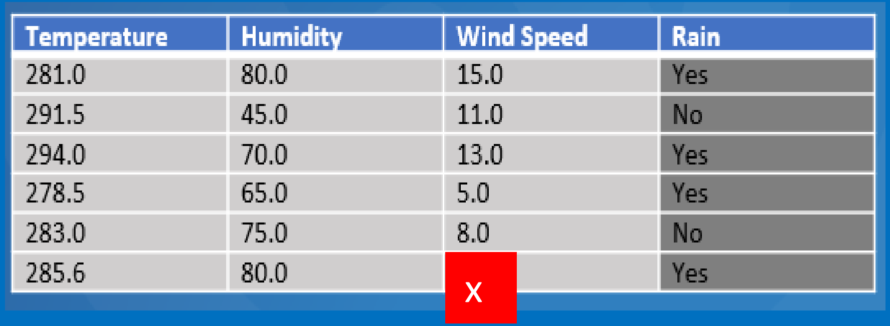
The weight for sample 2 $$w_2 = 11.0 \times \frac{0.3}{0.3+0.1} = 8.25$$

……

The weight for sample 5 $$w_5 = 8.0 \times \frac{0.1}{0.3+0.1} = 2.0$$

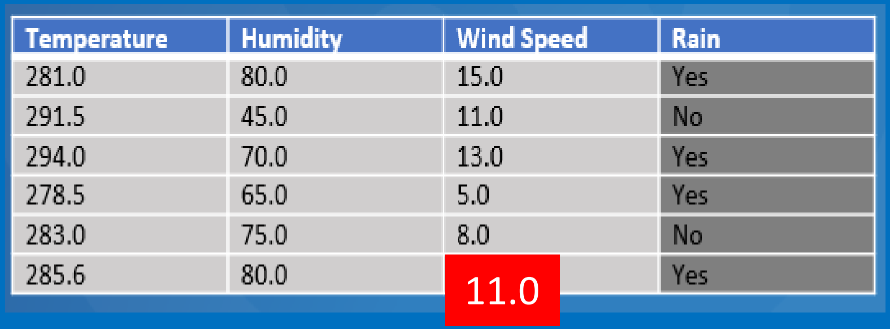
The weight for sample 6 $$w_6 = 0.0$$

Ultimately, the weighted guess for the missing data is 8.25 + 2.0 = 10.25



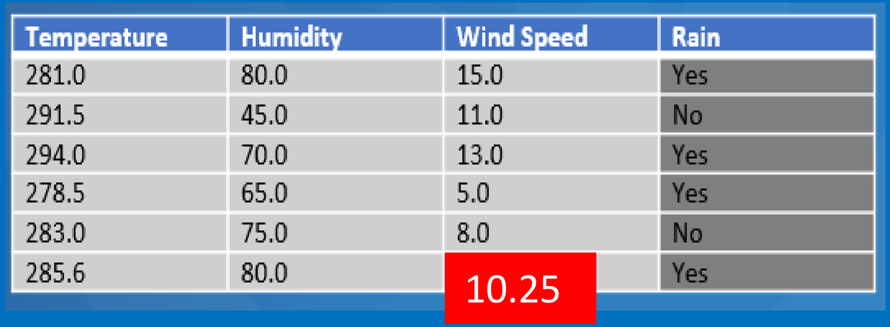
Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	x	Yes

Missing data



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

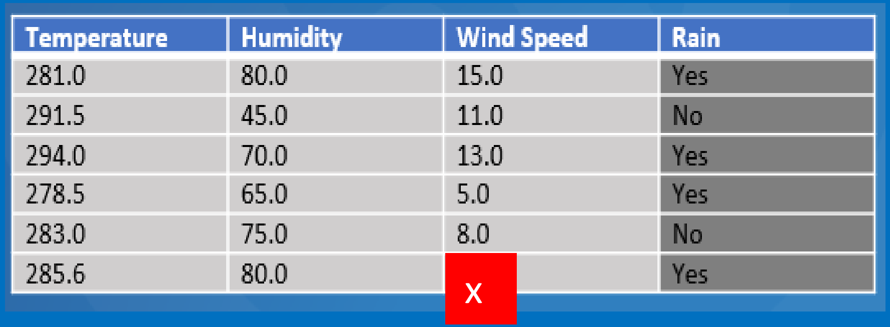
Simple average
(initial guess)



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	10.25	Yes

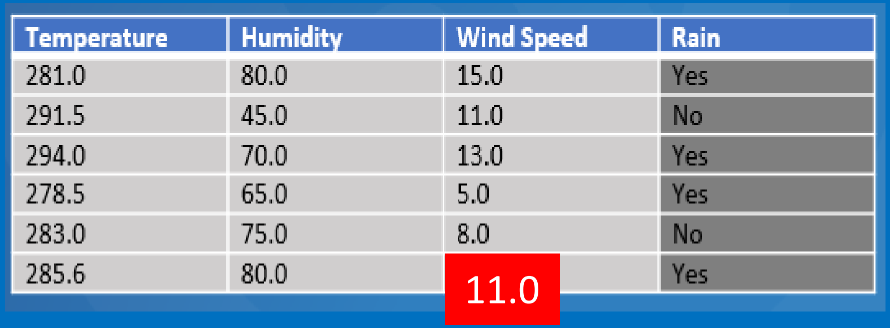
Weighted average
(2nd guess)

So we can see that how our guess for the missing data gets tuned



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	x	Yes

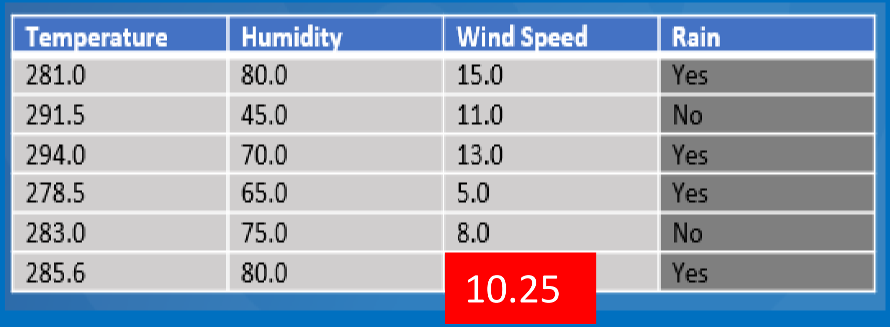
Missing data



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Simple average
(initial guess)

So we can see that how our guess for the missing data gets tuned



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	10.25	Yes

Weighted average
(2nd guess)

•••••••

Then we do this whole thing again

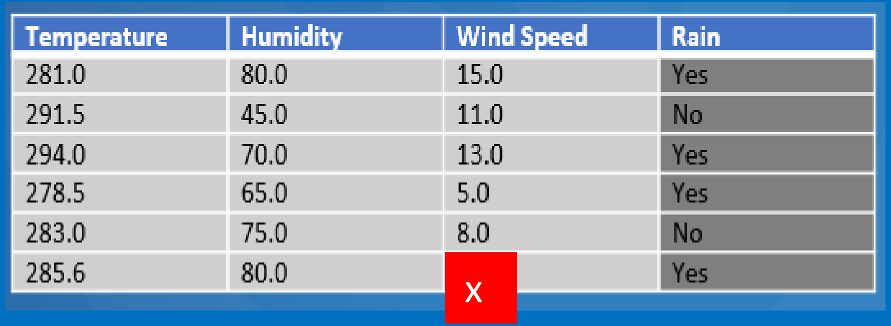
⇒ Revise our guess

⇒ Build a random forest,

⇒ Run the data through all the trees

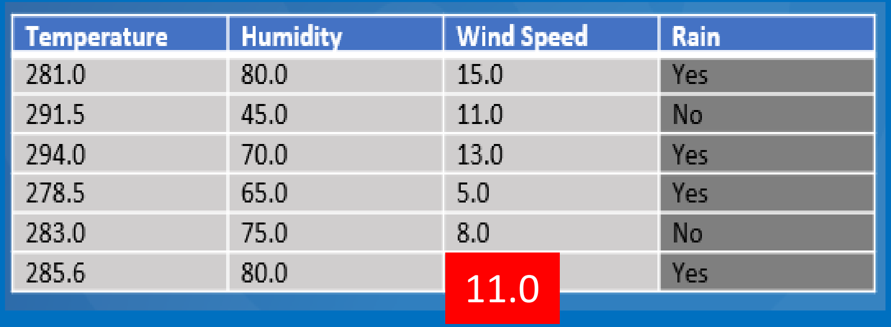
⇒ Recalculate the proximity matrix

⇒ Recalculate the missing data



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	x	Yes

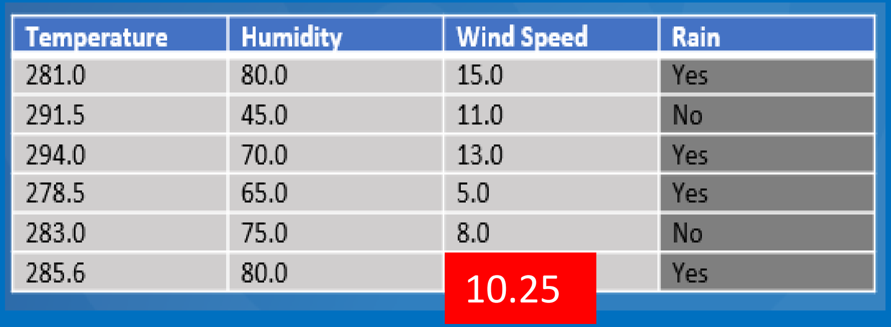
Missing data



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	11.0	Yes

Simple average
(initial guess)

So we can see that how our guess for the missing data gets tuned



Temperature	Humidity	Wind Speed	Rain
281.0	80.0	15.0	Yes
291.5	45.0	11.0	No
294.0	70.0	13.0	Yes
278.5	65.0	5.0	Yes
283.0	75.0	8.0	No
285.6	80.0	10.25	Yes

Weighted average
(2nd guess)

•••••••

Then we do this whole thing again
⇒ Revise our guess
⇒ Build a random forest,
⇒ Run the data through all the trees
⇒ Recalculate the proximity matrix
⇒ Recalculate the missing data

We do this many times, until the missing value does not change (converged)