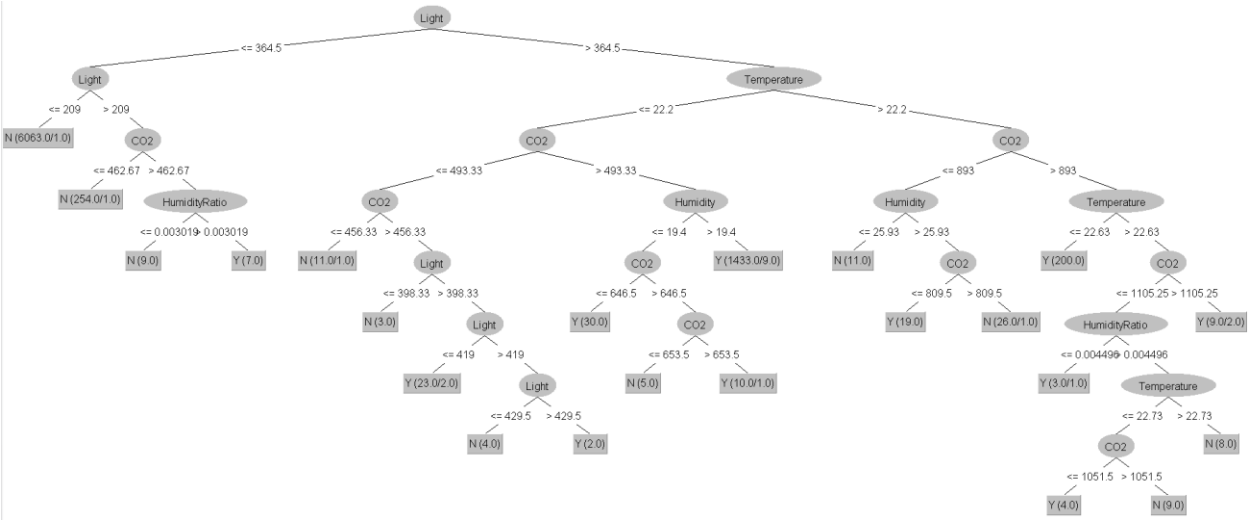
Following is the result given by Weka after running 10-fold cross-validation test using the **Decision Tree (J48) classifier**:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     occupancy_detection_training_dataset
Instances:    8143
Attributes:   7
              DateAndTime
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
              Occupancy
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
```

```
Light <= 364.5
|   Light <= 209: N (6063.0/1.0)
|   Light > 209
|   |   CO2 <= 462.67: N (254.0/1.0)
|   |   CO2 > 462.67
|   |   |   HumidityRatio <= 0.003019: N (9.0)
|   |   |   HumidityRatio > 0.003019: Y (7.0)
Light > 364.5
|   Temperature <= 22.2
|   |   CO2 <= 493.33
|   |   |   CO2 <= 456.33: N (11.0/1.0)
|   |   |   CO2 > 456.33
|   |   |   |   Light <= 398.33: N (3.0)
|   |   |   |   Light > 398.33
|   |   |   |   |   Light <= 419: Y (23.0/2.0)
|   |   |   |   |   Light > 419
|   |   |   |   |   |   Light <= 429.5: N (4.0)
|   |   |   |   |   |   Light > 429.5: Y (2.0)
|   |   CO2 > 493.33
|   |   |   Humidity <= 19.4
|   |   |   |   CO2 <= 646.5: Y (30.0)
|   |   |   |   CO2 > 646.5
|   |   |   |   |   CO2 <= 653.5: N (5.0)
|   |   |   |   |   CO2 > 653.5: Y (10.0/1.0)
|   |   |   Humidity > 19.4: Y (1433.0/9.0)
|   Temperature > 22.2
|   |   CO2 <= 893
|   |   |   Humidity <= 25.93: N (11.0)
|   |   |   Humidity > 25.93
|   |   |   |   CO2 <= 809.5: Y (19.0)
|   |   |   |   CO2 > 809.5: N (26.0/1.0)
|   |   CO2 > 893
|   |   |   Temperature <= 22.63: Y (200.0)
|   |   |   Temperature > 22.63
|   |   |   |   CO2 <= 1105.25
|   |   |   |   |   HumidityRatio <= 0.004496: Y (3.0/1.0)
|   |   |   |   |   HumidityRatio > 0.004496
|   |   |   |   |   |   Temperature <= 22.73
|   |   |   |   |   |   |   CO2 <= 1051.5: Y (4.0)
|   |   |   |   |   |   |   CO2 > 1051.5: N (9.0)
|   |   |   |   |   |   Temperature > 22.73: N (8.0)
|   |   |   |   CO2 > 1105.25: Y (9.0/2.0)

Number of Leaves  :     22

Size of the tree :      43
```

```
Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8092               99.3737 %
Incorrectly Classified Instances        51                0.6263 %
Kappa statistic                          0.9813
Mean absolute error                      0.0083
Root mean squared error                  0.0769
Relative absolute error                  2.4952 %
Root relative squared error             18.7977 %
Total Number of Instances             8143

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.005    0.983      0.988    0.985      0.981    0.993     0.977     Y
                 0.995    0.012    0.997      0.995    0.996      0.981    0.993     0.996     N
Weighted Avg.    0.994    0.011    0.994      0.994    0.994      0.981    0.993     0.992

=== Confusion Matrix ===

    a    b   <-- classified as
 1708   21 |    a = Y
   30 6384 |    b = N
```



How big was the tree?

Tree has 43 nodes in total with 22 leaf nodes. Size of the tree is reasonable as binary splitting of quantitative attributes have been used with appropriate threshold otherwise, size of the tree would have been much bigger. But if we increase the size of the tree to 55 nodes by unpruning it then there is a slight decrease in the training error, which can be neglected as there is a greater chance of increase in the test error due to overfitting. Additionally, upon increasing the size, ROC area also decreases significantly. If we decrease the size of the tree further then model becomes too much simplistic and it will not be true representative of the relationship between attributes and class values. So,

ideally the gap between training error and true error should be minimum and this size of the tree gives us that result.

What features did J48 choose as being important? Do these make sense?

Other than DateAndTime feature, all the other features were considered important by J48. Which is completely sensible, as there was no need of DateAndTime attribute for our use case. In the dataset DateAndTime is unique so there would have been many child nodes of this attribute equal to number of training cases and it would have been insensible to categorize this attribute based on threshold.

All of the chosen attributes contribute towards the decision of occupancy so they cannot be neglected. It was no surprise to see close relation between light and temperature and CO2 and humidity. So these choices made by J48 looks sensible.

Do you think that this decision tree corresponds with the way a human might perform the classification? Why or why not?

No, human might have performed differently. First of all it would have been difficult for the human to come up with the good threshold for each attribute that would minimize the loss. Even if he/she does come up with good threshold, it would have been very difficult for him/her to come up with these sequence of questions. For example, after first question of light, when the answer is lower than threshold then human would not have asked about light again he/she would have tried to jump to the conclusions after the negative answer to the first question as light is strong indicator of occupancy. Similarly, on level 2 human would not have asked question of CO2 level for both answers of temperature as this is counter intuitive.

Choice of confidence factor:

Following are the results given by Weka for different values of confidence factor:

Confidence Factor = 0.1

=== Run information ===

Number of Leaves: 19

Size of the tree: 37

Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8089	99.3369 %
Incorrectly Classified Instances	54	0.6631 %
Kappa statistic	0.9802	
Mean absolute error	0.0091	
Root mean squared error	0.0781	
Relative absolute error	2.7174 %	
Root relative squared error	19.0946 %	
Total Number of Instances	8143	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.987	0.005	0.982	0.987	0.984	0.980	0.993	0.976	Y
0.995	0.013	0.996	0.995	0.996	0.980	0.993	0.996	N
Weighted Avg:								
0.993	0.012	0.993	0.993	0.993	0.980	0.993	0.992	

=== Confusion Matrix ===

a	b	<-- classified as
1706	23	a = Y
31	6383	b = N

Confidence Factor = 0.15

=== Run information ===

Number of Leaves: 22

Size of the tree: 43

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8091	99.3614 %
Incorrectly Classified Instances	52	0.6386 %
Kappa statistic	0.9809	
Mean absolute error	0.0085	
Root mean squared error	0.0766	
Relative absolute error	2.5374 %	
Root relative squared error	18.724 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.982	0.988	0.985	0.981	0.994	0.982	Y
0.995	0.012	0.997	0.995	0.996	0.981	0.994	0.996	N
Weighted Avg:								
0.994	0.011	0.994	0.994	0.994	0.981	0.994	0.993	

=== Confusion Matrix ===

a	b	<-- classified as
1708	21	a = Y
31	6383	b = N

Confidence Factor = 0.2

=== Run information ===

Number of Leaves: 22

Size of the tree: 43

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.0084	
Root mean squared error	0.0759	
Relative absolute error	2.5103 %	
Root relative squared error	18.5559 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.986	0.982	0.994	0.982	Y
0.995	0.012	0.997	0.995	0.996	0.982	0.994	0.996	N
Weighted Avg:								
0.994	0.011	0.994	0.994	0.994	0.982	0.994	0.993	

=== Confusion Matrix ===

a	b	<-- classified as

1708 21 | a = Y

29 6385 | b = N

Confidence Factor = 0.25

=== Run information ===

Number of Leaves: 22

Size of the tree: 43

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8092	99.3737 %
Incorrectly Classified Instances	51	0.6263 %
Kappa statistic	0.9813	
Mean absolute error	0.0083	
Root mean squared error	0.0769	
Relative absolute error	2.4952 %	
Root relative squared error	18.7977 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.985	0.981	0.993	0.977	Y
0.995	0.012	0.997	0.995	0.996	0.981	0.993	0.996	N

Weighted Avg:

0.994	0.011	0.994	0.994	0.994	0.981	0.993	0.992

=== Confusion Matrix ===

a	b	<-- classified as
1708	21	a = Y
30	6384	b = N

Confidence Factor = 0.3

=== Run information ===

Number of Leaves: 22

Size of the tree: 43

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.0082	
Root mean squared error	0.0767	
Relative absolute error	2.4604 %	
Root relative squared error	18.7603 %	

Total Number of Instances 8143

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.986	0.982	0.992	0.977	Y
0.995	0.012	0.997	0.995	0.996	0.982	0.992	0.995	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.992	0.991	

=== Confusion Matrix ===

a	b	<-- classified as
1709	20	a = Y
30	6384	b = N

Confidence Factor = 0.35

=== Run information ===

Number of Leaves: 23

Size of the tree: 45

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8093 99.386 %

Incorrectly Classified Instances 50 0.614 %

Kappa statistic 0.9817

Mean absolute error 0.008

Root mean squared error 0.0768

Relative absolute error 2.4022 %

Root relative squared error 18.7795 %

Total Number of Instances 8143

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.986	0.982	0.991	0.973	Y
0.995	0.012	0.997	0.995	0.996	0.982	0.991	0.994	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.991	0.990	

=== Confusion Matrix ===

a	b	<-- classified as
1709	20	a = Y
30	6384	b = N

Confidence Factor = 0.4

Number of Leaves: 23

Size of the tree: 45

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.008	
Root mean squared error	0.0768	
Relative absolute error	2.4022 %	
Root relative squared error	18.7795 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.986	0.982	0.991	0.973	Y
0.995	0.012	0.997	0.995	0.996	0.982	0.991	0.994	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.991	0.990	

=== Confusion Matrix ===

a	b	<-- classified as
1709	20	a = Y
30	6384	b = N

Confidence Factor = 0.45

Number of Leaves: 23

Size of the tree: 45

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.008	
Root mean squared error	0.0768	
Relative absolute error	2.4022 %	
Root relative squared error	18.7795 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.986	0.982	0.991	0.973	Y
0.995	0.012	0.997	0.995	0.996	0.982	0.991	0.994	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.991	0.990	

=== Confusion Matrix ===

a	b	<-- classified as
1709	20	a = Y
30	6384	b = N

Confidence Factor = 0.5

Number of Leaves: 23

Size of the tree: 45

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.008	
Root mean squared error	0.0768	
Relative absolute error	2.4022 %	
Root relative squared error	18.7795 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.005	0.983	0.988	0.986	0.982	0.991	0.973	Y
0.995	0.012	0.997	0.995	0.996	0.982	0.991	0.994	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.991	0.990	

=== Confusion Matrix ===

a	b	<-- classified as
1709	20	a = Y
30	6384	b = N

Confidence Factor = 0.6

Number of Leaves: 25

Size of the tree: 49

Time taken to build model: 0.54 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.008	
Root mean squared error	0.0769	
Relative absolute error	2.3959 %	
Root relative squared error	18.7986 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.989	0.005	0.982	0.989	0.986	0.982	0.991	0.971	Y
0.995	0.011	0.997	0.995	0.996	0.982	0.991	0.994	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.991	0.989	

=== Confusion Matrix ===

a	b	<-- classified as
1710	19	a = Y
31	6383	b = N

Confidence Factor = 0.7

Number of Leaves: 25

Size of the tree: 49

Time taken to build model: 0.46 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8093	99.386 %
Incorrectly Classified Instances	50	0.614 %
Kappa statistic	0.9817	
Mean absolute error	0.008	
Root mean squared error	0.0769	
Relative absolute error	2.3959 %	
Root relative squared error	18.7986 %	
Total Number of Instances	8143	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.989	0.005	0.982	0.989	0.986	0.982	0.991	0.971	Y
0.995	0.011	0.997	0.995	0.996	0.982	0.991	0.994	N
Weighted Avg:								
0.994	0.010	0.994	0.994	0.994	0.982	0.991	0.989	

=== Confusion Matrix ===

a	b	<-- classified as
1710	19	a = Y
31	6383	b = N

Analysis:

99.386% is the highest accuracy achieved by decision tree model and highest ROC area is 0.994. As we go from 0.1 confidence factor to 0.2 accuracy decreases slightly then increases and ROC area increases significantly. At 0.2 both accuracy and ROC area have highest values. When confidence factor = 0.25 accuracy slightly decreases and ROC area decrease as well. At 0.3 accuracy again goes to highest value and ROC area drops down. From 0.35 to 0.7, values of both accuracy and ROC area becomes consistent they do not change. But, time taken to build the tree increases significantly for confidence factor > 0.6 because large number of branches and nodes needs to be

computed. If confidence factor is higher then it becomes computationally expensive. So confidence factor of 0.2 gives the best result where there is a nice trade-off between sensitivity and specificity and also accuracy is good as well.

Comparison of the performance of J48 with pruning and without pruning:

ZeroR as Testbase:

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -V -result-matrix "weka.experiment.ResultMatrixPla
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/4/19 12:32 PM


Dataset                   (1) rules.ZeroR '' | (2) trees.J48 ' (3) trees.J48 ' (4) bayes.Naive
----------------------------------------------------------------------------------------------
occupancy_detection_train(100)   78.77(0.04) |   99.32(0.28) v   99.33(0.27) v   97.70(0.55) v
----------------------------------------------------------------------------------------------
                                       (v/ /*) |         (1/0/0)         (1/0/0)         (1/0/0)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-U -M 2' -217733168393644444
(3) trees.J48 '-C 0.2 -M 2' -217733168393644444
(4) bayes.NaiveBayes '' 5995231201785697655
```

As we can see that accuracy of J48 without pruning, J48 with pruning (confidence factor: 0.2) and Naive Bayes is greater than ZeroR and there is ‘v’ attached next to their results, which indicates that difference in the accuracy of their results as compared to ZeroR is statistically significant. So we can say that these two algorithms achieved a statistically significantly better result than the ZeroR baseline. Since, the accuracy for J48 with pruning is highest, so next we want to see if the accuracies of other algorithms are significantly lower than J48 with pruning or not. For this purpose we make J48 with pruning a Test base.

J48 with pruning (Confidence Factor: 2) as Test Base:

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -V -result-matrix "weka.experiment.ResultMatrixPla
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/4/19 12:35 PM


Dataset                   (3) trees.J48 '-C  | (1) rules.ZeroR (2) trees.J48 ' (4) bayes.Naive
-------------------------------------------------------------------------------------------------
occupancy_detection_train(100)   99.33(0.27) |   78.77(0.04) *   99.32(0.28)     97.70(0.55) *
-------------------------------------------------------------------------------------------------
                                      (v/ /*) |         (0/0/1)         (0/1/0)         (0/0/1)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-U -M 2' -217733168393644444
(3) trees.J48 '-C 0.2 -M 2' -217733168393644444
(4) bayes.NaiveBayes '' 5995231201785697655
```

In this case accuracy score of both J48 and Naive Bayes is lower than that of J48 with pruning and also there is a ‘*’ next to the results of Naive Bayes which means that its accuracy is significantly lower. So Naive Bayes can be ruled out of the competition.

Accuracy score and also the standard deviation of both J48 with pruning and without pruning are very close to each other and there is no significant difference between them, so they both can be a good model. But, since we need to minimize true error (keeping overfitting in mind) as well as computation so, J48 with pruning would be a good choice.