Imperial College London

Introduction to ML
Decision Tree Coursework 1

17/10/2022 - 04/11/2022

MSc Computing in Artificial Intelligence
MSc Computing in Artificial Intelligence & Machine Learning

Students :

Pierre-Antoine ARSAGUET
pea22@ic.ac.uk || AI & ML

Louis BERTHIER
ldb22@ic.ac.uk || AI & ML

Nikita DMITRIEFF
nnd19@ic.ac.uk || AI

Ryan EL KHOURY
re122@ic.ac.uk || AI

Contents

List of figures

Chapter A

Presentation of the coursework

This project aims at answering the following question: In which room is a person located according to the intensities of the WIFI signals collected from his mobile phone?

A case study is shown in figure 1 below.

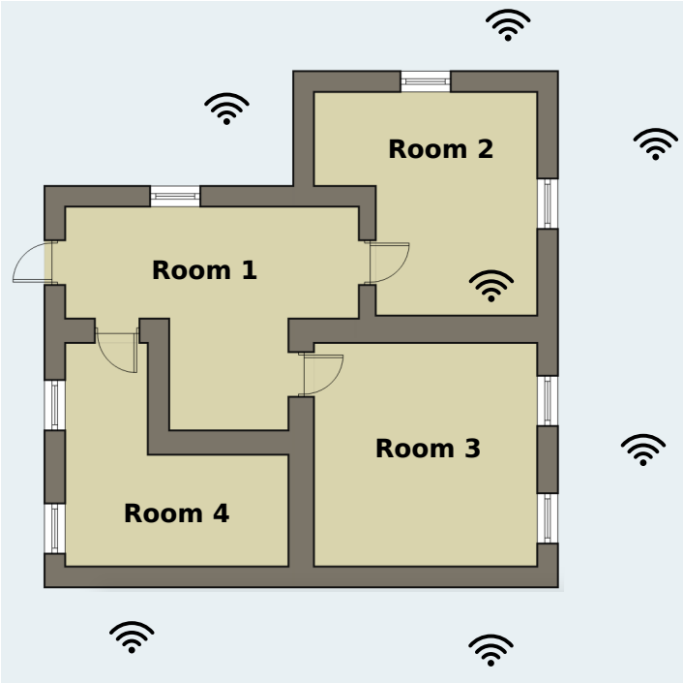

Figure 1: Case study of our project with 7 emitters and 4 rooms

A decision tree algorithm, implemented from scratch, will be used to classify a user's location based on the strength of the WIFI signals relative to the 7 emitters.

Two sets of data were made available, namely:

- One identified as **clean**
- One identified as **noisy**

The aim of this report is to present the findings of the algorithm implemented, along with complementary methods improving its performance, such as cross-validation and/or pruning. A succinct analysis of the results will follow, based on multiple metrics.

Chapter B

Step 3: Evaluation without pruning

B.1 Clean dataset

Results associated with the balanced clean dataset before pruning are presented in this section.
Results were obtained from cross-validation repeated **10 times** (outer) and **5 times** (inner) with a depth determined automatically via the tree construction.
Here, no model comparison via the use of a validation dataset was performed. For each fold, a model is trained on the training dataset and directly tested on the test dataset.
General parameters for clean and noisy datasets before pruning are presented below.

- train proportion = 0.8
- folds number = 10
- seed = 60012

B.1.1 Confusion Matrix

The confusion matrix of each split was normalized. The global confusion matrix could then be created by averaging over these normalized matrices. Hence, the global confusion matrix of the clean dataset without pruning is presented below.

Global Confusion Matrix (Clean Unupruned)				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	0.97343479	0	0.01469665	0.01186855
Room 2	0.	0.91835862	0.08164138	0.
Room 3	0.00816481	0.05873228	0.92565716	0.00744575
Room 4	0.01020603	0.	0.00290545	0.98688852

Additional confusion matrices for each fold from another simulation can be found in the appendix .1.

B.1.2 Final results

Final results for the clean dataset, before pruning, are presented below.

Clean Dataset Unpruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.95175
Accuracy (Std)	-	-	-	-	0.0109
Precision (Mean)	0.98234	0.93744	0.90483	0.98088	0.95137
Precision (Std)	0.01177	0.02322	0.02614	0.01363	0.01869
Recall (Mean)	0.97343	0.91836	0.92566	0.98689	0.95109
Recall (Std)	0.02193	0.01967	0.02751	0.01231	0.02036
F1 (Mean)	0.9777	0.92767	0.91484	0.98377	0.95010
F1 (Std)	0.01209	0.01844	0.02165	0.00811	0.01507

The results matrix associated with the 10 folds from appendix .1 can be found in appendix .2.

B.2 Noisy dataset

Results associated with the slightly imbalanced dataset are presented in this section

As with the clean dataset, 10 models were studied with cross-validation, keeping the same general parameters.

B.2.1 Confusion Matrix

The confusion matrix of each split was normalized. The global confusion matrix could then be created by averaging over these normalized matrices. Hence, the global confusion matrix of the noisy dataset without pruning is presented below.

Global Confusion Matrix (Noisy Unpruned)				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	0.70256378	0.04298627	0.07227361	0.18217633
Room 2	0.0546237	0.80726651	0.08803161	0.05007818
Room 3	0.05061221	0.09659228	0.78759609	0.06519941
Room 4	0.14343854	0.04464423	0.06974505	0.74217218

Additional confusion matrices for each fold from another simulation can be found in the appendix .3.

B.2.2 Final results

Final results for the noisy dataset, before pruning, are presented below.

Noisy Dataset Unpruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.75975
Accuracy (Std)	-	-	-	-	0.03388
Precision (Mean)	0.74158	0.79949	0.78365	0.71507	0.75995
Precision (Std)	0.06979	0.03515	0.05414	0.0569	0.0541
Recall (Mean)	0.70256	0.80727	0.7876	0.74217	0.7599
Recall (Std)	0.07194	0.04683	0.0431	0.06038	0.05556
F1 (Mean)	0.72067	0.80255	0.78412	0.7275	0.75871
F1 (Std)	0.06606	0.03323	0.03524	0.05273	0.04682

The results matrix associated with the 10 folds from appendix .3 can be found in appendix .4.

B.3 Result analysis before pruning

For the clean dataset, results show a greater performance for identifying rooms 1 and 4. Indeed, at least 5% difference in the metrics can be seen compared to rooms 2 and 3. However, the noisy dataset shows an opposite trend. The model performs better when identifying rooms 2 and 3 with regard to rooms 1 and 4, showing a clear 5% difference in metrics validation there as well.

B.4 Dataset differences

Unsurprisingly, real gain in performance can be seen across our various metrics for the clean dataset compared to the noisy dataset, with accuracy improving by as much as 20%. The noisy dataset is not perfectly balanced with differences in the number of observations of up to 3%.
When training on the noisy dataset, the decision tree model will tend to be deeper than in the case of the clean dataset. This results in possible over-fitting problems with the model. The model becomes too specific to the train set and performs worse for the test set.

Chapter C

Step 4: Evaluation with pruning

C.1 Clean dataset after pruning

Results associated with the clean dataset after pruning was performed are presented in this section. Results were obtained from cross-validation repeated **10 times** (outer) and **5 times** (inner) with a depth determined automatically via the tree construction.
A model comparison was performed via the use of a validation dataset (inner). For each fold, a model, trained on a train dataset, is pruned using a validation dataset. The best model is selected according to its accuracy performance. The best model is then evaluated on the test dataset.
General parameters for clean and noisy datasets with pruning are presented below.

- train proportion = 0.8
- validation proportion = 0.2
- folds number (outer) = 10
- folds number (inner) = 5
- seed = 60012

C.1.1 Confusion Matrix

The confusion matrix of each split was normalized. The global confusion matrix could then be created by averaging over these normalized matrices. Hence, the global confusion matrix of the clean dataset after pruning is presented below.

Global Confusion Matrix (Clean Pruned)				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	0.99112828	0.	0.00783005	0.00104167
Room 2	0.	0.90161052	0.09838948	0.
Room 3	0.01244526	0.04376019	0.93580904	0.00798551
Room 4	0.01274657	0.	0.01070465	0.97654879

Additional confusion matrices for each fold from another simulation can be found in the appendix .5.

C.1.2 Final results

Final results for the clean dataset, after pruning was performed, are presented below.

Clean Dataset Pruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.95175
Accuracy (Std)	-	-	-	-	0.00807
Precision (Mean)	0.97573	0.95504	0.88972	0.99084	0.95283
Precision (Std)	0.0184	0.02879	0.03686	0.01375	0.02445
Recall (Mean)	0.99113	0.90161	0.93581	0.97655	0.95128
Recall (Std)	0.01031	0.03381	0.03289	0.01093	0.02199
F1 (Mean)	0.98321	0.92681	0.9112	0.9835	0.95118
F1 (Std)	0.00807	0.01759	0.01929	0.0046	0.01239

The results matrix associated with the 10 folds from appendix .5 can be found in appendix .6.

C.2 Noisy dataset after pruning

Results associated with the noisy dataset after pruning are presented in this section

C.2.1 Confusion Matrix

The confusion matrix of each split was normalized. The global confusion matrix could then be created by averaging over these normalized matrices. Hence, the global confusion matrix of the noisy dataset after pruning is presented below.

Global Confusion Matrix (Noisy Pruned)				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	0.72945486	0.03403832	0.05311178	0.18339505
Room 2	0.0396947	0.84496727	0.09041264	0.02492538
Room 3	0.03956045	0.08036949	0.83281928	0.04725078
Room 4	0.16703859	0.03075137	0.03988489	0.76232515

Additional confusion matrices for each fold from another simulation can be found in the appendix .7.

C.2.2 Final results

Final results for the noisy dataset, after pruning was performed, are presented below.

Noisy Dataset Pruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.7945
Accuracy (Std)	-	-	-	-	0.05025
Precision (Mean)	0.73941	0.85333	0.83254	0.75672	0.7955
Precision (Std)	0.11837	0.03354	0.05327	0.10191	0.07677
Recall (Mean)	0.72945	0.84497	0.83282	0.76233	0.79239
Recall (Std)	0.11222	0.03813	0.04594	0.12077	0.07927
F1 (Mean)	0.73137	0.84842	0.83128	0.75487	0.79158
F1 (Std)	0.10964	0.02671	0.03521	0.10326	0.06871

The results matrix associated with the 10 folds from appendix .7 can be found in appendix .8.

C.3 Result analysis after pruning

Compared to the performance without pruning, similar results can be found for the clean dataset, while a performance gain is observed for the noisy dataset of about 5% on all metrics.
For the clean dataset, the model was already very efficient but the pruning allows us to have a simpler model which remains just as efficient. For the noisy dataset, pruning allowed us to remove 'unnecessary' nodes and branches which reduced the over-fitting of our model.

C.4 Depth analysis

In the case of the clean dataset, the tree depth remains similar on average with or without pruning (or slightly lower with pruning) whereas, for the noisy dataset, a lower average depth can be seen when using pruning.
It is therefore estimated that the deeper a tree is, the better it will perform on the train dataset. However, for accuracy, the opposite effect occurs. Indeed, a great depth indicates that our model is very probably over-fitting (problems of memorization and not generalization occur) and thus that its accuracy on unknown data will not be very high.

Chapter D

Visualization of trees

A method to visualize a tree from scratch using Numpy and Matplotlib was also implemented. Visualizations of the trained tree on the entire clean dataset are shown in figure 2 and 3.

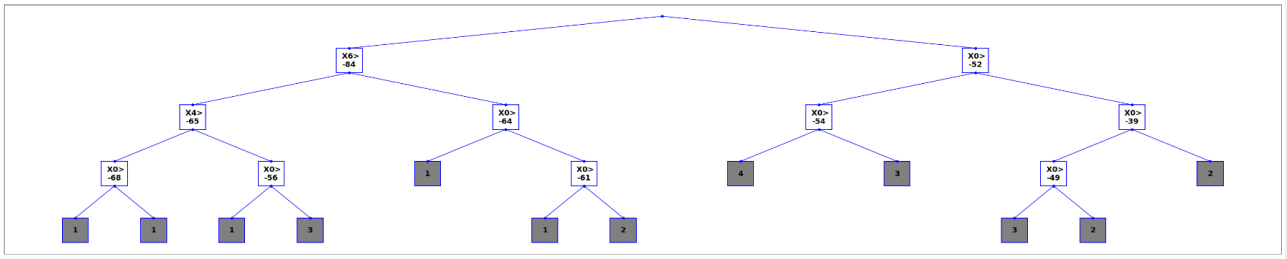

Figure 2: Visualization of our tree without pruning

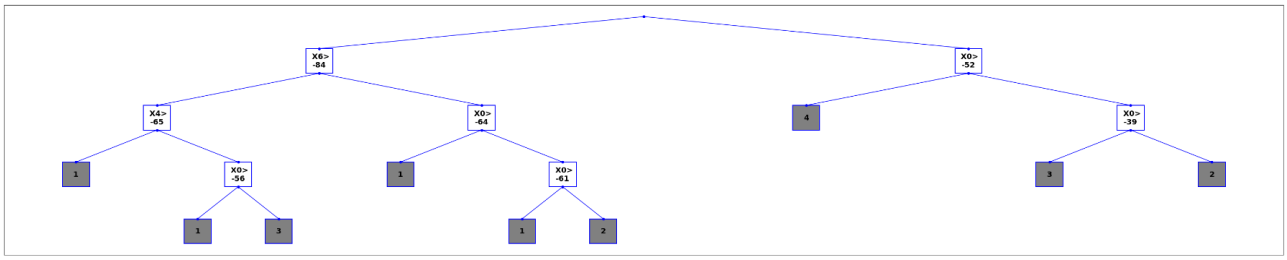

Figure 3: Visualization of our tree after pruning

A depth is specified directly in the function (here the depth is 5) in order to have a correct viewing window at the output. Beyond a certain depth, it is not easy to read the graph correctly or to upload it correctly to this report.

Appendices

.1 Other matrices: Clean and Unpruned

Confusion Matrix fold 1				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	96	0	0	1
Room 2	0	95	11	0
Room 3	1	10	98	0
Room 4	0	0	0	88

Confusion Matrix fold 2				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	102	0	3	0
Room 2	0	80	8	0
Room 3	0	12	95	1
Room 4	1	0	0	98

Confusion Matrix fold 3				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	103	0	0	0
Room 2	0	96	7	0
Room 3	2	4	80	0
Room 4	5	0	0	103

Confusion Matrix fold 4				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	104	0	0	1
Room 2	0	84	11	0
Room 3	1	4	96	2
Room 4	0	0	0	97

Confusion Matrix fold 5				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	76	0	1	1
Room 2	0	101	11	0
Room 3	1	6	96	0
Room 4	2	0	1	104

Confusion Matrix fold 6				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	104	0	2	0
Room 2	0	92	7	0
Room 3	1	6	92	1
Room 4	2	0	0	93

Confusion Matrix fold 7				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	105	0	1	0
Room 2	0	82	6	0
Room 3	0	6	106	1
Room 4	1	0	1	91

Confusion Matrix fold 8				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	103	0	0	0
Room 2	0	84	10	0
Room 3	1	7	93	0
Room 4	0	0	0	102

Confusion Matrix fold 9				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	110	0	0	0
Room 2	0	94	3	0
Room 3	1	6	88	0
Room 4	0	0	1	97

Confusion Matrix fold 10				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	98	0	2	1
Room 2	0	104	10	0
Room 3	3	8	85	1
Room 4	0	0	1	87

.2 Results: Clean and Unpruned

Clean Dataset Unpruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.95050
Accuracy (Std)	-	-	-	-	0.01077
Precision (Mean)	0.97830	0.92950	0.90603	0.98952	0.95084
Precision (Std)	0.01734	0.02493	0.02385	0.00925	0.01997
Recall (Mean)	0.98679	0.91594	0.91545	0.98514	0.95083
Recall (Std)	0.01120	0.02381	0.02129	0.01401	0.01758
F1 (Mean)	0.98234	0.92242	0.91058	0.98723	0.95064
F1 (Std)	0.01067	0.01903	0.01975	0.00693	0.01410

.3 Other matrices: Noisy and Unpruned

Confusion Matrix fold 1				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	60	12	6	17
Room 2	3	75	7	7
Room 3	4	7	80	9
Room 4	14	11	6	82

Confusion Matrix fold 2				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	59	7	17	17
Room 2	3	80	8	10
Room 3	10	8	76	5
Room 4	11	2	7	80

Confusion Matrix fold 3				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	67	4	7	21
Room 2	5	87	12	6
Room 3	2	2	82	11
Room 4	18	5	4	67

Confusion Matrix fold 4				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	76	6	4	12
Room 2	7	84	6	3
Room 3	4	9	86	4
Room 4	17	7	6	69

Confusion Matrix fold 5				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	61	3	2	21
Room 2	1	85	13	8
Room 3	5	9	84	7
Room 4	16	4	4	77

Confusion Matrix fold 6				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	65	6	5	21
Room 2	7	88	10	2
Room 3	8	6	84	6
Room 4	19	10	2	61

Confusion Matrix fold 7				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	66	9	8	28
Room 2	4	77	14	4
Room 3	7	11	76	5
Room 4	14	4	10	63

Confusion Matrix fold 8				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	71	3	7	19
Room 2	5	84	7	3
Room 3	9	9	83	15
Room 4	17	7	2	59

Confusion Matrix fold 9				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	57	0	7	17
Room 2	9	86	10	2
Room 3	3	17	88	4
Room 4	13	7	3	77

Confusion Matrix fold 10				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	74	4	7	28
Room 2	10	70	3	6
Room 3	6	5	81	5
Room 4	24	1	7	69

.4 Results: Noisy and Unpruned

Noisy Dataset Unpruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.74900
Accuracy (Std)	-	-	-	-	0.02169
Precision (Mean)	0.70677	0.80955	0.79665	0.68622	0.74980
Precision (Std)	0.03072	0.04914	0.04936	0.05572	0.04624
Recall (Mean)	0.67083	0.80715	0.79596	0.72004	0.74850
Recall (Std)	0.05354	0.02245	0.0372	0.04165	0.03871
F1 (Mean)	0.68694	0.80727	0.79536	0.70172	0.74782
F1 (Std)	0.03251	0.02406	0.03502	0.04143	0.03326

.5 Other matrices: Clean and Pruned

Confusion Matrix fold 1				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	101	0	0	0
Room 2	0	91	6	1
Room 3	1	5	85	3
Room 4	1	0	0	106

Confusion Matrix fold 2				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	107	0	1	0
Room 2	0	84	8	0
Room 3	2	3	91	0
Room 4	3	0	1	100

Confusion Matrix fold 3				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	104	0	1	0
Room 2	0	83	12	0
Room 3	0	7	82	0
Room 4	0	0	1	110

Confusion Matrix fold 4				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	95	0	1	3
Room 2	0	87	8	0
Room 3	1	4	96	4
Room 4	2	0	1	98

Confusion Matrix fold 5				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	92	0	0	0
Room 2	0	97	14	0
Room 3	3	2	85	0
Room 4	1	0	4	102

Confusion Matrix fold 6				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	105	0	0	0
Room 2	0	95	11	0
Room 3	3	1	91	0
Room 4	2	0	0	92

Confusion Matrix fold 7				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	124	0	0	0
Room 2	0	81	10	0
Room 3	2	3	100	0
Room 4	0	0	0	80

Confusion Matrix fold 8				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	97	0	0	0
Room 2	0	91	18	0
Room 3	1	4	86	1
Room 4	0	0	1	101

Confusion Matrix fold 9				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	101	0	0	0
Room 2	0	93	11	0
Room 3	2	3	85	0
Room 4	1	0	2	102

Confusion Matrix fold 10				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	91	0	0	0
Room 2	0	95	8	0
Room 3	0	3	93	1
Room 4	3	0	0	106

.6 Results: Clean and Pruned

Clean Dataset Pruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.95125
Accuracy (Std)	-	-	-	-	0.00896
Precision (Mean)	0.97314	0.96214	0.88286	0.98778	0.95148
Precision (Std)	0.01449	0.01740	0.03760	0.02111	0.02265
Recall (Mean)	0.99408	0.89427	0.93805	0.97796	0.95109
Recall (Std)	0.01208	0.02678	0.01790	0.01406	0.01771
F1 (Mean)	0.98341	0.92670	0.90909	0.98268	0.95047
F1 (Std)	0.00966	0.01722	0.02128	0.01276	0.01523

.7 Other matrices: Noisy and Pruned

Confusion Matrix fold 1				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	81	4	7	10
Room 2	2	72	7	2
Room 3	0	11	94	0
Room 4	12	2	21	75

Confusion Matrix fold 2				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	84	3	7	10
Room 2	5	70	9	4
Room 3	7	5	94	4
Room 4	15	1	9	73

Confusion Matrix fold 3				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	71	7	9	18
Room 2	3	87	4	1
Room 3	4	9	95	4
Room 4	18	1	6	63

Confusion Matrix fold 4				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	83	2	5	7
Room 2	2	95	4	5
Room 3	6	8	75	2
Room 4	5	3	9	89

Confusion Matrix fold 5				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	94	4	1	4
Room 2	4	78	9	3
Room 3	6	7	89	4
Room 4	12	6	0	79

Confusion Matrix fold 6				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	64	2	2	22
Room 2	5	91	6	2
Room 3	2	13	87	2
Room 4	17	1	9	74

Confusion Matrix fold 7				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	87	4	2	21
Room 2	5	88	4	5
Room 3	7	8	76	5
Room 4	12	3	1	72

Confusion Matrix fold 8				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	80	6	5	6
Room 2	4	85	3	3
Room 3	3	6	95	2
Room 4	4	3	4	91

Confusion Matrix fold 9				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	83	2	5	5
Room 2	2	79	5	4
Room 3	3	10	93	2
Room 4	10	4	2	91

Confusion Matrix fold 10				
Actual \ Prediction	Room 1	Room 2	Room 3	Room 4
Room 1	93	2	8	13
Room 2	4	86	8	2
Room 3	5	11	60	7
Room 4	14	3	4	80

.8 Results: Noisy and Pruned

Noisy Dataset Pruned Results					
	Room 1	Room 2	Room 3	Room 4	Average
Accuracy (Mean)	-	-	-	-	0.82400
Accuracy (Std)	-	-	-	-	0.03208
Precision (Mean)	0.80546	0.84618	0.83331	0.81463	0.82490
Precision (Std)	0.05248	0.02267	0.06109	0.06888	0.05128
Recall (Mean)	0.80207	0.86750	0.83623	0.78822	0.823505
Recall (Std)	0.06817	0.03279	0.04819	0.06359	0.05319
F1 (Mean)	0.80301	0.85616	0.83317	0.79960	0.82300
F1 (Std)	0.05539	0.01810	0.04168	0.05702	0.04305