CONCORDIA UNIVERSITY

Department of Building Civil & Environmental Engineering

Classification of Building Design Baseline (DT & RF)

Course: Project and Report III

Instructor: Dr. Mazdak Nik-bakht

Student's Name: Seyed Mahyar Mousavi Mohammadi

Student ID: 40084040

Winter and Summer 2020

Abstract:

This report explains the procedure of training a classifier to predict whether a combination of building design specifications has acceptable Energy Usage Intensity (EUI) or Benefit to Cost Ratio (BCR). Decision Tree (DT) and Random Forest (RF) are the classifiers which are trained and evaluated separately for both EUI and BCR classifications. For each model the parameters are optimized, and the performances are compared to decide the more reliable one. The results demonstrates that the RF model classifies slightly better in comparison with the DT model with an accuracy of 96.47% for the EUI Classification and an accuracy of 96.53% for the BCR classification. One advantage of this project is that by using only one optimized model and only ten attributes out of 50, excluding the target attributes, the generated RF model can classify the target values for both EUI and BCR with a high reliability.

TABLE OF CONTENT:

TABLE OF FIGURES:

1. Introduction

1.1 Motivation and background

The increasing energy consumption worldwide has been worrying due to the limitations of energy resources as well as the environmental issues such as climate change, global warming, and etc. The statistics announced by the International Energy Agency shows that during the years 1984 to 2004, primary energy consumption and the emissions have increased by 49% and 43% respectively and this trends are expected to grow annually by almost 2%[1].The buildings consume a large portion of the worldwide energy and since the demand is increasing constantly with the growing living standards, designing energy efficient buildings should be considered as a mandate. In order to design buildings which are efficient in terms of energy performances, the designers must find the influential building design factors and their combinations that lead to energy efficient products. Energy simulation softwares such as OpenStudio, TRACE 700, etc. are being used to predict the energy performance of different building designs and operations. Recently, many studies have tried to apply data science to generate models that can predict the performance of the buildings using different combinations of design factors[2]. In this study, Decision Tree and Random Forest have been used to classify the design factors combinations into binary values which define whether the designs are better than a certain baseline. The baselines for EUI and BCR are defined according to the codes and standards related to the energy efficiency and life cycle cost aspects respectively.

1.2 Problem Statement

This project aims to classify the effective parameters of building design on the energy usage of the building and their benefit to cost ratios (BCR). For this purpose, the Energy Usage Intensity (EUI) and Benefit Cost Ratio (BCR) considered as binary targets which say whether the EUI and BCR are better or worse than defined baselines.

1.3 Objectives

The objective is to create a reliable model among different methods in such a way that can predict which combinations result in EUI and BCR better or worse than the baseline i.e. the final model is going to be capable of distinguishing the design combinations resulting in EUI and BCR better or worse than the baselines with acceptable performances.

2. Literature Review:

Researchers have conducted several studies on the application of data analysis specially for data classification and regression. Zhun Yu et al (2010) used decision tree to predict the buildings energy demand and reached the accuracy of 93% for the training set and 92% for the test set[2]. Yildiz et al (2017) investigated regression models and ANN models to estimate the daily peak electricity load. The highest performance between regression models was for SVR (Support Vector Regression) and the highest performance between ANN models was for BR (Backpropagation)[3]. Kim et al (2010) conducted a study on the annual energy cost of an energy efficient building design through the application of Decision Tree[4]. The study demonstrated the HVAC system resulted in having the most impact on the annual energy cost and the building orientation selected as the least effective parameter in the annual energy cost. Xiao and Fan (2014) have used Entropy weighted k-means (EWKM) clustering, association rule mining, and decision tree to improve building operational performance. The decision tree is used to detect the association rules abnormalities[5]. Copozzoli et al (2015) implemented regression and classification trees to classify primary energy demand into three classes and for the evaluation of the most influencing factors on the classification. U-value was concluded to be the most influencing factor and the first node of split in the decision tree[6]. Ashari et al (2013) investigated to compare the performance of Naive Bayes (NB), Decision Tree (DT), and K-Nearest Neighbour (KNN) on finding alternative designs in an energy simulation tool. The NB method generated the highest performance measures among the three methods[7]. Yang et al (2017) have used KNN and NB classification methods for appliance disaggregation to achieve energy efficiency in residential buildings and KNN resulted in higher performance compared to the NB[8].

3. Methodology:

The methodology for both EUI and BCR classifications are the same since the dataset attributes are common between these two procedures.

The CRISP-DM procedure is chosen for implementing the project, so each step of CRISP-DM is explained below to have a better resolution of work:

3.1 Understanding the dataset:

The dataset for data mining includes 57 attributes and 14688 datapoints. Below are the list of attributes, their spread, and their related categories:

Table 1 - List of all attributes

Main Attributes	**Type**	**Spread**		Category
		Min, Least	**Max, Most**	
Data_point_ID	Integer	1	14688	ID
Building_model	Polynominal	17 types		General Information
Multiuse_Building? (1: yes; 0: no)	Binominal	0	0	
Main_Building_Type	Polynominal	12 types		
Rise_(low_or_high)	Binominal	High (5184)	Low (9504)	
Stories	Integer	1	13	
Data_center_in_the_Building? (1: yes; 0: no)	Binominal	1(864)	0(13824)	
Height_(m)	Real	3.05	51.48	Architectural Parameters
Building_Volume_(m^3)	Real	708.37	126016.35	
Roof_footprint_(f^2)	Integer	2786	126672	
Roof_footprint_(m^2)	Real	258.83	11768.21	
Orientation	Polynominal	90(1584)	0(7344)	
Wall_Area_(f^2)	Real	1721	74849	
Wall_Area_(m^2)	Real	159.89	6953.7	
South_Wall_area(m^2)	Real	46.47	3476.86	
North_Wall_area(m^2)	Real	46.47	3476.86	
East_Wall_area(m^2)	Real	46.47	2317.91	
West_Wall_area(m^2)	Real	46.47	2317.91	
South_Wall_%	Real	16.1776	57.3789	
North_Wall_%	Real	16.1776	56.8327	
East_Wall_%	Real	7.6564	35.5911	
West_Wall_%	Real	7.6564	35.5911	
WWR_South	Real	0.0019	0.991	
WWR_North	Real	0	0.991	
WWR_East	Real	0	0.991	
WWR_West	Real	0	0.991	
Stories/Roof_area	Real	0.0001	0.013	
Stories/Wall_area	Real	0.0004	0.006	
Height/Roof_area	Real	0.001	0.039	
Height/Wall_area	Real	0.002	0.0191	
Volume/Roof_area	Real	2.604	35.367	
Volume/Wall_area	Real	4.4305	24.5455	
Volume/Thermal_Zone	Real	98.3239	13747.9052	
Wall_area/Roof_area	Real	0.2376	3.4513	
South_Wall/Volume	Real	0.0099	0.0747	
North_Wall/Volume	Real	0.0099	0.0701	
East_Wall/Volume	Real	0.0087	0.0656	
West_Wall/Volume	Real	0.0087	0.0656	
U-value_(W/(m2K))	Real	1.624	3.122	Energy efficiency
R-value_Roof	Integer	21	65	
R-value_Wall	Integer	16	41	
Heating %	Integer	24	64	
SHGC	Real	0.476	0.762	
Number_of_glazings(0: dbl; 1: trp)	Binominal	6 Triple types (5328)	6 Double types (9360)	Window type
Air_or_Arg_(0: air; 1: arg)	Binominal	1(4752)	0(9936)	
Lighting_Type	Polynominal	Troffer (4320)	Surface_Ambient(5472)	Lighting power
LPD_Reduction	Integer	1	38	
Thermal_Zones	Integer	2	118	HVAC system design
HVAC_System	Polynominal	11 types of HVAC Systems		
Heating_thermostat_Sch(hr/week)	Integer	72	168	
Cooling_thermostat_Sch(hr/week)	Integer	60	168	
Heating_Setpoint_Temp	Integer	15.6	21.7	
Cooling_Setpoint_Temp	Integer	22.2	29.44	
EUI	Real	11.8	213.5	Target attributes
EUI Classification (0: better than the baseline; 1: worse than the baseline)	Binominal	0(6608)	1(8080)	
NPW	Real	18417.7	25994729.92	
NWP Classification (0: worse than the baseline; 1: better than the baseline)	Binominal	1(3937)	0(10751)	

The target attribute for this project is either EUI (Energy Usage Intensity) classification or BCR (Benefit to Cost Ratio) classification depending on the aim of the modelling. Also, the EUI values are turned into binominal format to compare the different designs of buildings to the baseline.

The dataset shows different factors affecting the target attribute in different designs for different building categories. The purpose will be to classify the target values based on various design factors. In addition, it could be found which attributes have more influences on the labeled attribute in terms of being better or worse than the specified baseline.

3.2 Data preparation:

To prepare the dataset for training the decision tree, there are some points to consider.

- **Missing/Inconsistent values**: there is neither missing nor inconsistent values in the dataset.
- **Outliers**; based on the statistics of the dataset, there are some values to be considered as outliers. Since the Decision tree is not sensitive to outliers, they are ignored.
- **Attribute Selection**: Among all the attributes in the initial dataset, the ID related attributes, which do not affect the results, are excluded in the first place. The exact amounts of EUI are also set aside since it is only needed to recognize which class each design combination belongs.

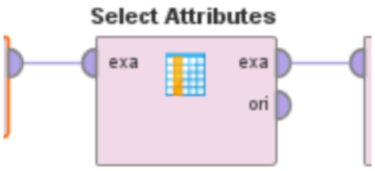

Figure 1 - Select Attributes operator in RapidMiner

- **Defining Target Attribute**: Labelling an attribute ensures that the class distribution in the sample output from stratified sampling is the same as in the whole ExampleSet. It also helps having a target attribute for classification[9].
  After importing the dataset using "read CSV" parameter into the RapidMiner (RM), first step was setting the target role through the "Set Role" parameter in RM.

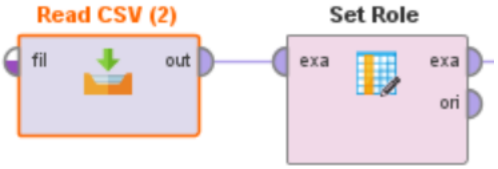

Figure 2 - Labelling the target attribute in RapidMiner

- **Discretize:** In order to feed the decision tree with the proper data, continuous numerical datapoints must be categorized into several groups. The Discretize operator is used to fit a range of numbers into different bins[9]. In this study all the numerical values from each attribute are placed into bins.

Figure 3 - Discretize operator in RapidMiner

3.3 Modeling:

3.3.1 Decision Tree:

The decision tree consists of a root and some nodes which are the most influential attributes on defining the target values. The tree starts splitting from the root using a large portion of the prepared dataset called the training set and continues splitting at each node based on the selected criterion such as information gain and gain ratio. It continues growing till it meets the stopping factors like maximal depth and minimal leaf size (the last node of each tree is a leaf). At this point the model is completed and ready to be measured for its goodness.

Also, Bagging can be used to Improve the reliability of the DT. This operator, AKA bootstrap aggregating, is a meta-algorithm in machine learning. It combines multiple models generated from different portions of the training set and uses them for voting. This procedure results in having a more stable and more accurate final classification and regression model as well as reducing variance and avoiding overfitting. The learner, i.e. DT places inside this nested operator and then the created model, will be optimized. In this operator, each time, 70% of the training set is used randomly to come up with ten different models for voting[9].

Figure 4 - Decision Tree operator in RapidMiner

3.3.2 Random Forest:

Random forest is made up of multiple uncorrelated random trees with specific parameters in which each tree is trained by a bootstrapped sample of the whole dataset and delivers a class prediction. This classifier follows a simple and fundamental rule of thumb which is the wisdom of crowds. The random trees are trained by the bootstrapped samples from the dataset. Only one sample, among all the samples, is investigated to choose the rule which dictates whether the tree should split at each node (attribute). After training multiple trees, the random forest model is applied to an unseen datapoints. Each tree produces a class prediction and the resulting model votes according to the results from all trained random trees. All the class predictions have the same importance, so the final class prediction has lower variance comparing the prediction from each random tree[9].

Figure 5 - Random Forest operator in RapidMiner

3.4 Evaluation:

The measures of goodness to decide whether the classifier is accurate enough are **accuracy**, **precision**, **recall**, **F measures,** and **AUC** (the higher the AUC, the better the model classifies the positive and negative classes). Cross-validation is used to evaluate the classifier. The former randomly splits the data, trains the model using the training set using DT in Bagging operator and applies the model on the test set and does the evaluation of the model using appropriate performance operator. The latter partitions the dataset into “n” equal subsets, including n-1 subsets as the training set and one subset as a test set. Then it repeats this loop n times and gives the average of the results as the output. In this study, ten folds are applied (i.e. n=10).

The central part of the evaluation is choosing the proper performance operator that produces the aimed parameters (accuracy, precision, recall, and f-measures)[9]. Regarding the type of target values, which are Boolean, the Binominal Performance operator is used.

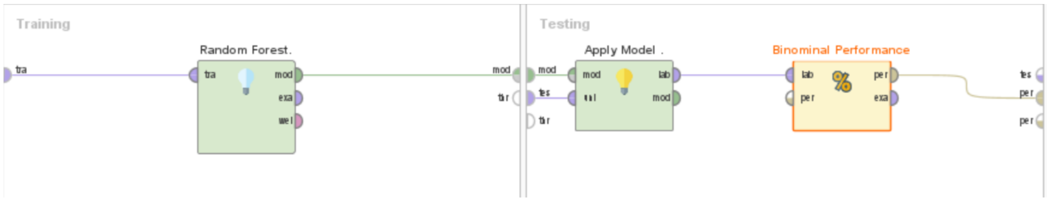

Figure 6 - Cross Validation operator environment in RapidMiner

4. Implementation and Results

In this section methods are investigated for two separate target features which are Energy Usage intensity (EUI) and Benefit to Cost Ratio (BCR). Figure 7 demonstrates the procedure of Decision Tree and Random Forest modeling process. In this process attributes are selected as the first process on the prepared dataset, then the target attribute is selected, and the numerical values are categorized into bins, then the parameters of the model is selected, and the model performance is evaluated. If the performance is not good enough, the number of bins as well as the model parameters must be adjusted till an acceptable performance is reached.

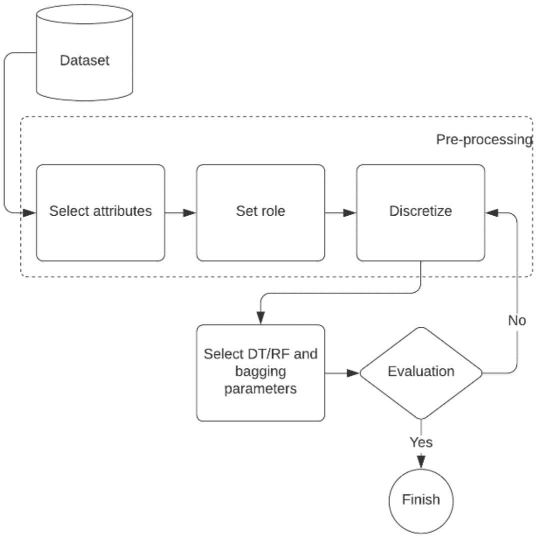

Figure 7 - DT and RF implementation process

4.1 EUI:

Two implemented methods on EUI are Decision Tree (DT) and Random Forest (RF).

4.1.1 Decision Tree:

The procedure of training of a decision tree includes the following steps:

4.1.1.1 Attribute selection:

For selecting the proper attributes to start the modelling, the “Select Attributes” parameter was used. Between existing attributes, having tried various sets for better model accuracy, three groups of attributes were chosen as table below:

Table 2 - Three attribute selections

Group 1	Group 2		Group 3
All Attributes	East_Wall/Volume	South_Wall_%	HVAC_System
	East_Wall_%	Stories	Lighting_Type
	EUI Classification	Stories/Wall_area	LPD_Reduction
	Height/Roof_area	U-value_(W/(m2K))	Main_Building_Type
	Height/Wall_area	Volume/Roof_area	North_Wall/Volume
	Height_(m)	Volume/Wall_area	Orientation
	HVAC_System	Wall_area/Roof_area	R-value_Roof
	LPD_Reduction	West_Wall/Volume	SHGC
	North_Wall/Volume	West_Wall_%	U-value_(W/(m2K))
	North_Wall_%	WWR_East	Volume/Thermal_Zone
	Orientation	WWR_North	WWR_East
	R-value_Roof	WWR_South	WWR_North
	SHGC	WWR_West	WWR_South
	South_Wall/Volume		WWR_West

Group 1 attributes include all the attributes.

Group 2 attributes are those attributes which were supposed to be investigated through the project.

Group 3 attributes are selected after removing multiple attributes which did not have any contribution to the model accuracy of the project. These attributes were attained through the processes of trial and error during the project.

EUI Classification attribute is included in all three groups.

4.1.1.2 Discretizing:

Before the modelling process, the numerical attributes with continuous values should be discretized. At the primary step, 3 bins were chosen those attributes, but after trying different numbers of bins, the best results were for 5 bins.

As there were not a lot of different values in some categorical attributes, they stayed without any change.

4.1.1.3 DT parameters selection:

The next step is choosing Decision Tree parameters to start the modelling process. There were three different criterions was used in this project:

- Gain Ratio
- Information Gain
- Gini Index

For each of the above criterions, several related values should be selected which are:

- Maximal depth limits the depth of the tree.

- Confidence is used for error calculation of pruning.
- Minimal gain determines if the data should split at one node or not.
- Minimal leaf size; defines the minimum number of examples in the leaves.
- Minimal size for split; sets the least amount of datapoints at each node to be splittable.
- Number of pre pruning alternatives; adjusts the number of alternative nodes tested for splitting.

The following amounts were assigned to DT parameters to train the first Decision Tree using group 1 attributes:

Table 3 - The parameters of the first optimized DT (trained by group 1 attributes)

Parameter	**Initial Selection**
Criterion	Gain Ratio
Maximal depth	10
Confidence	0.1
Minimal gain	0.02
Minimal leaf size	2
Minimal size for split	4
Number of pre pruning alternatives	3

Pre-pruning and post pruning were also applied to reduce the size of the tree, and its complexity as well as to avoid overfitting.

4.1.1.4 Evaluation:

The assessment of the concluded model generated the following results:

Table 4 Evaluation of the First Optimized DT (trained by group 1 attributes)

Class	Accuracy	Precision	Recall	F Measure
0	92.10%	96.20%	92.88%	94.51%
1		82.00%	90.25%	85.93%

AUC - ROC curve is a measure of goodness for classification at different thresholds. ROC is a probability curve and AUC shows the degree of separability. It demonstrate the capability of model in classification. The AUC and ROC are plotted in the figure below:

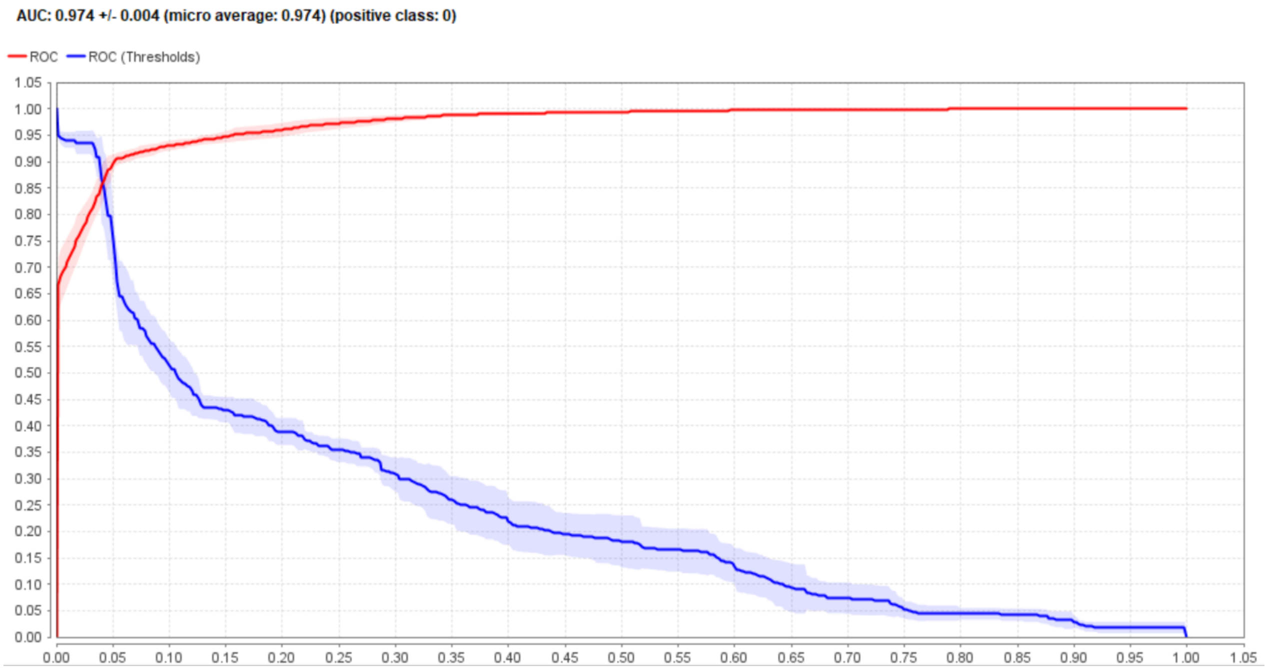

Figure 8 - Area Under the Curve (AUC) and receiver operating characteristic curve (ROC) for the DT benchmark for EUI Classification

Selecting gain ratio criterion in the trained DT, the following attributes are recognized as the most important ones to define the target values.

Table 5 - List of effective attributes

Attribute	**Weight**
Orientation	0.327
R-value_Roof	0.148
LPD_Reduction	0.122
Thermal_Zones	0.058
Heating %	0.046
Stories	0.041
Height/Wall_area	0.041
WWR_West	0.039
WWR_South	0.028
U-value_(W/(m2K))	0.027
WWR_North	0.025
R-value_Wall	0.022
East_Wall/Volume	0.018
SHGC	0.014
Heating_thermostat_Sch(hr/week)	0.009
Cooling_thermostat_Sch(hr/week)	0.009
Lighting_Type	0.007
WWR_East	0.007
Air_or_Arg_(0: air; 1: arg)	0.004
South_Wall_%	0.003
Roof_footprint_(m^2)	0.002
HVAC_System	0.002

Trying different DT parameter combinations on different attribute selections and evaluating their performances resulted in a model and the measures of goodness are shown in the tables below:

Table 6 - performance measures for different groups of attributes

Attribute Selection	**Performance Type**	**Positive Class**	**Accuracy**	**Precision**	**Recall**	**F measure**
Group 1	Cross Validation	1	92.93%	85.60%	88.49%	87.02%
		0		95.74%	94.55%	95.14%
Group 2	Cross Validation	1	75.35%	93.41%	8.64%	15.81%
		0		74.89%	99.78%	85.56%
Group 3	Cross Validation	1	92.73%	86.84%	85.98%	86.41%
		0		94.87%	95.22%	95.04%

The AUC and ROC metrics are plotted separately for each group in the following figures:

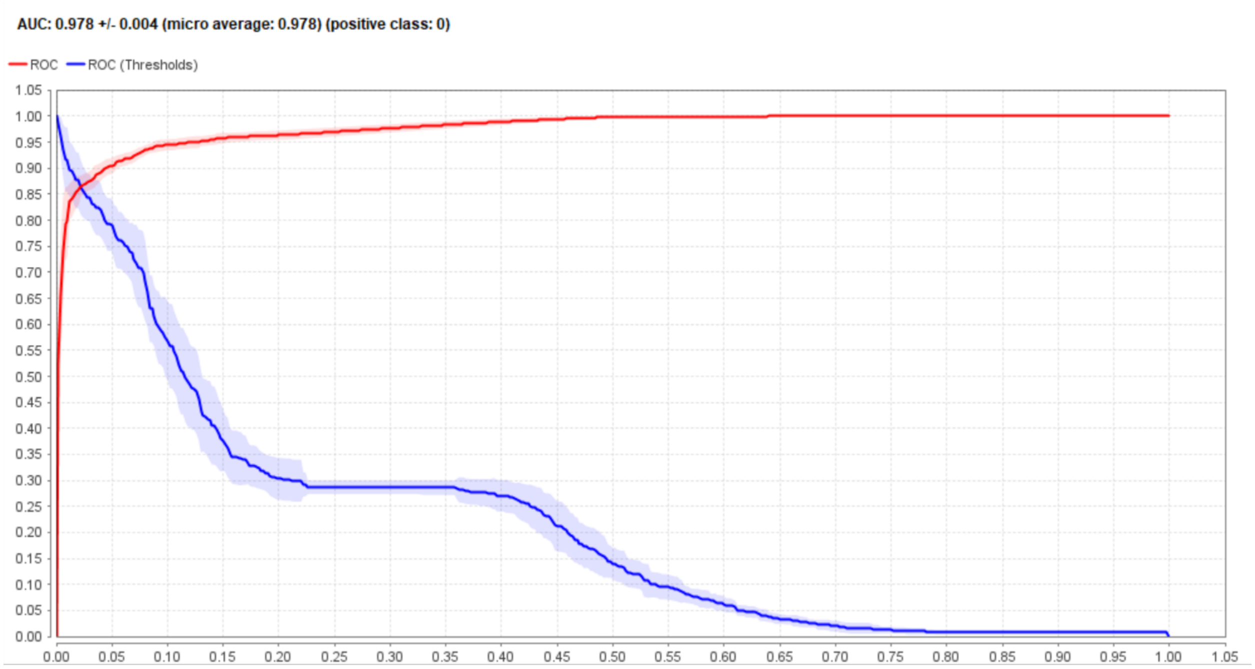

Figure 9 – AUC and ROC for the first optimized DT for EUI Classification (Group 1)

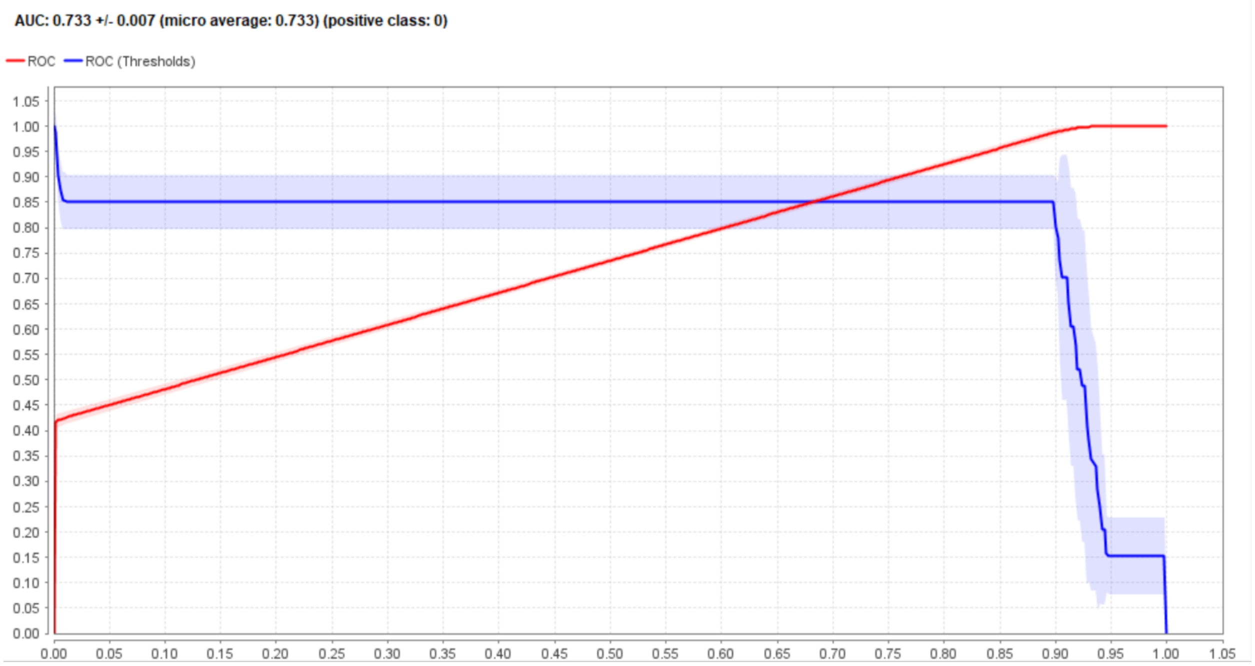

Figure 10 - AUC and ROC for the first optimized DT for EUI Classification (Group 2)

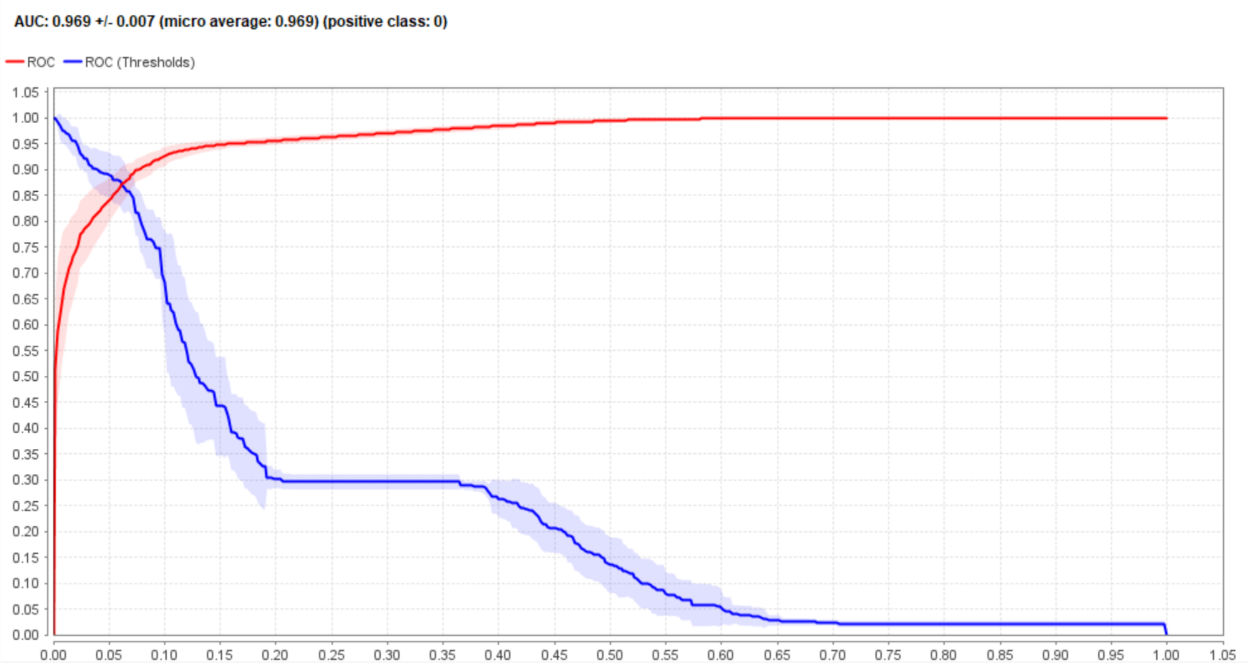

Figure 11 - AUC and ROC for the first optimized DT for EUI Classification (Group 3)

The selected parameters for the final model are:

Table 7 - Final model Parameters

Operator	Parameter	Selection
Decision Tree	Criterion	Gini Index
	maximal depth	10
	confidence	0.4
	minimal gain	0.1
	minimal leaf size	5
	minimal size for split	5
	Number of pre pruning alternatives	5
Bagging	Sample ratio	0.8
	iterations	10

4.1.1.5 Optimization

The Optimize operators are new features recently added to RM, which are very useful to find the best model based on several named parameters as well as the influencing attributes on the target values.

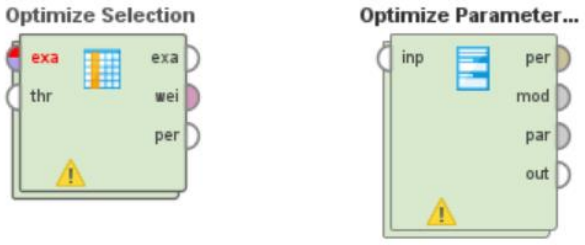

Figure 12 - Optimization operators in RapidMiner

Optimize parameter: The operator runs a process with different predefined combinations of parameters to generate the combination with the highest accuracy[9]. Selecting all attributes, this operator was experimented with several combinations of parameter values to find the most accurate class predictions among them. The best obtained combinations in terms of accuracy are shown below:

Table 8 - Best combinations including 3 investigated DT criterion

Operator	Parameter	Criterion		
		gini_index	information_gain	gain_ratio
Decision Tree	Maximal Depth	10	10	10
	Confidence	0.46	0.42	0.42
	Minimal Gain	0.12	0.1	0.06
	Minimal Leaf Size	3	3	3
	Prepruning Alternatives	10	3	3
	Minimal Size for Split	6	3	4
Discretize	bins	5	5	3
Bagging	sample ratio	0.8	0.8	0.8
	iteration	10	10	10
Performance (Cross Validation)	Accuracy	95.17%	**96.41%**	91.15%

As table 8 shows, the information gain criterion generates results with higher accuracy.

Optimize selection: This operator receives a dataset and uses the model and the evaluation to define the most relevant attributes to the classification of target values[9]. The application of this operator on the above models with above specifications introduced three sets of attributes based on the selected criterion:

Table 9 - The most influential attributes on EUI classification based on the selected criterion

#	Attribute Name	Criterion		
		Gini Index	**Information Gain**	**Gain Ratio**
1	Heating %	✗	✗	✓
2	Heating_thermostat_Sch(hr/week)	✗	✗	✓
3	HVAC_System	✓	✓	✓
4	Lighting_Type	✓	✓	✓
5	LPD_Reduction	✓	✓	✓
6	Main_Building_Type	✓	✓	✗
7	North_Wall/Volume	✓	✓	✗
8	Number_of_glazings(0: dbl; 1: trp)	✗	✗	✓
9	Orientation	✓	✓	✓
10	R-value_Roof	✗	✓	✓
11	R-value_Wall	✓	✗	✓
12	SHGC	✓	✓	✗
13	Stories/Roof_area	✗	✗	✓
14	Thermal_Zones	✗	✗	✓
15	U-value_(W/(m2K))	✓	✓	✗
16	Volume/Thermal_Zone	✗	✓	✓
17	WWR_East	✓	✓	✓
18	WWR_North	✓	✓	✓
19	WWR_South	✓	✓	✓
20	WWR_West	✓	✓	✓

Three groups of attributes (group 1, group 2, and group 4 which is the result of Optimized Selection operator) are investigated to find out how the best modelling procedure generated from the optimization (table 8) works. Table 10 shows the performance of the model on the group generated from using information gain as the criterion.

Table 10 - Final Optimization (evaluation results)

Attribute Selection	Performance Type	Positive Class	Accuracy	Precision	Recall	F measure
Group 1	Cross Validation	1	96.41%	94.99%	91.41%	93.17%
		0		96.90%	98.23%	97.56%
Group 2	Cross Validation	1	90.48%	76.65%	92.71%	83.92%
		0		97.11%	89.66%	93.23%
Group 4	Cross Validation	1	96.47%	95.05%	91.62%	93.30%
		0		96.97%	98.25%	97.61%

Group 4 includes attributes from table 9 which are generated using information gain criterion in the DT.

As It's shown in table 9, the attributes generated from Optimize Selection based on information gain (Group 4), are the same as group 3 attributes which are generated via trial and error.

AUC and ROC are illustrated in the figures below:

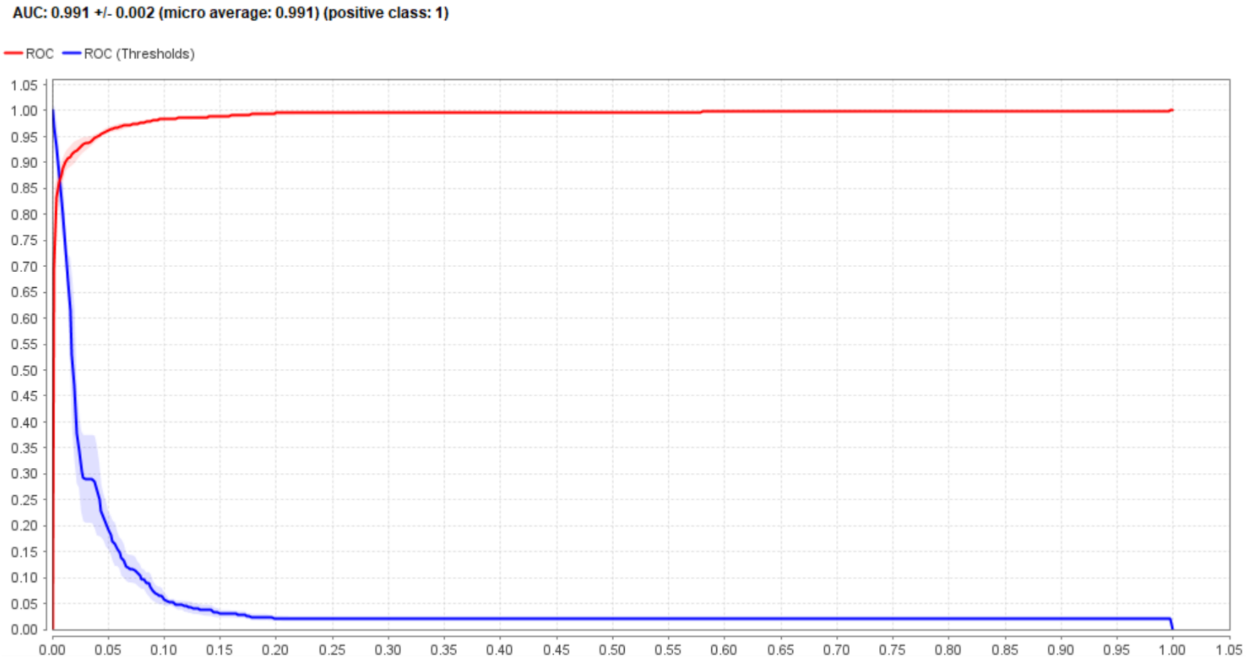

Figure 13 - AUC and ROC for the final optimized DT for EUI Classification (Group 1)

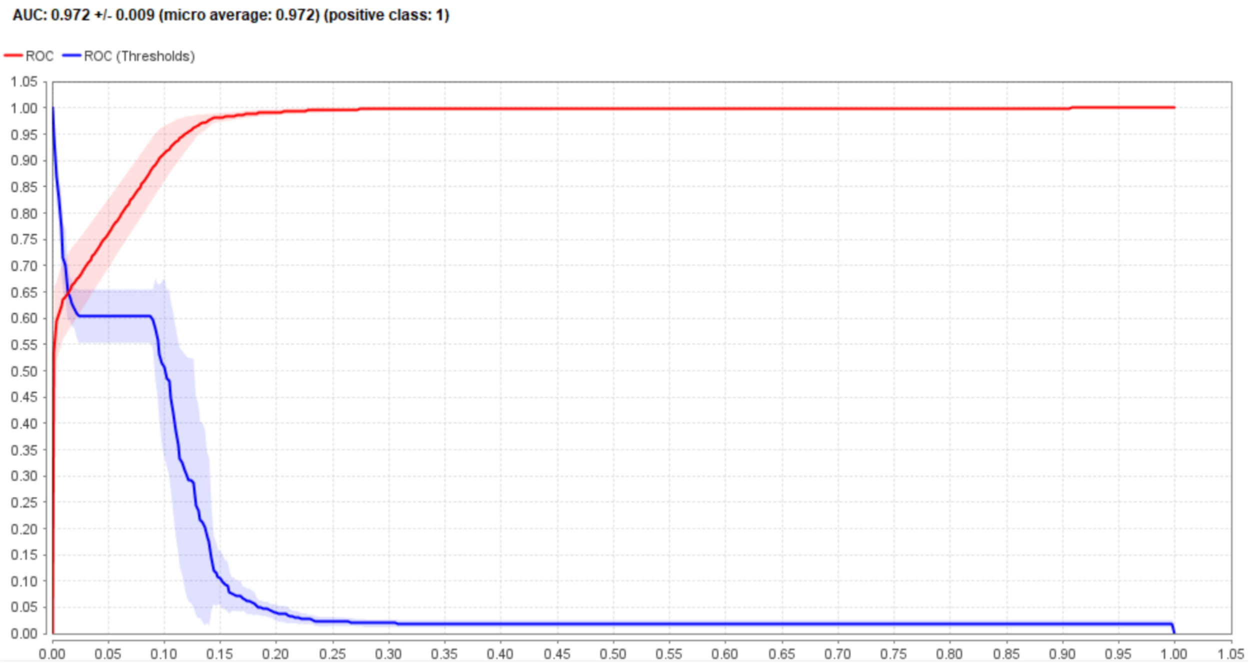

Figure 14 - AUC and ROC for the final optimized DT for EUI Classification (Group 2)

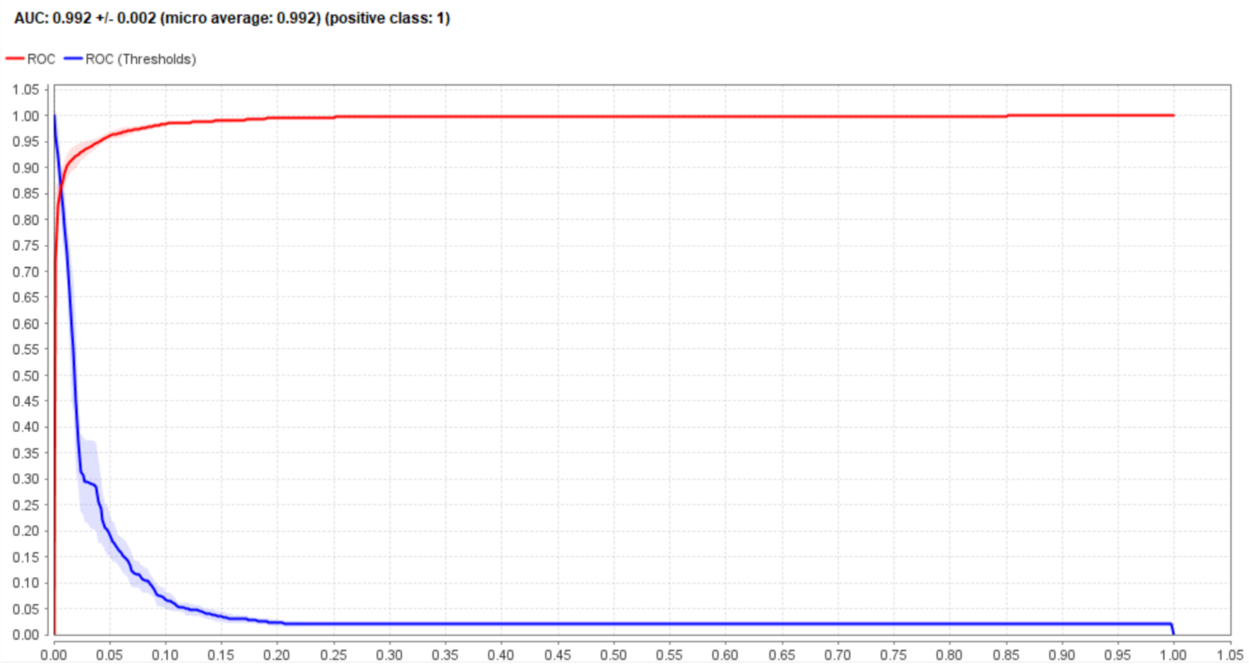

Figure 15 - AUC and ROC for the final optimized DT for EUI Classification (Group 4)

4.1.1.6 Random Forest

Since the decision tree forms the basis of random forest, All the steps are in the same order, and the only difference would be the model parameters. To avoid creating so many models which are not desired in terms of accuracy, a benchmark model was first created by the RM default parameter values for Random Forest operator with the following values and performance:

Table 11 – Selected parameters values and properties for the RF benchmark model

Selected Attributes	Group 1
Discretize: Number of bins	3
Random Forest Criterion	Gain_ratio
Random Forest Number of Trees	100
Maximal Depth	10
Random Forest Voting Strategy	confidence vote
Pruning confidence	0.1
Prepruning minimal gain	0.01
Prepruning minimal leaf size	2
Prepruning minimal size for split	4
Number of prepruning alternatives	3
Accuracy	89.99%

Table 12 - Performance of the RF benchmark model

accuracy: 89.99%			
	true 1	true 0	class precision
pred. 1	2688	222	92.37%
pred. 0	1249	10529	89.40%
class recall	68.28%	97.94%	

4.1.1.7 Optimization

The first step in optimization is to find the most important attributes in determining the target values classes. By the application of Optimize Selection operator, the following attributes are recognized as the most determinative features in this classification which are listed separately based on the selected criterion:

Table 13 - determinative attributes on the target values based on the selected criterion

#	Attribute	Criterion		
		Gini Index	Information Gain	Gain Ratio
1	Height_(m)	✓	✗	✗
2	HVAC_System	✗	✓	✓
3	Lighting_Type	✓	✓	✓
4	LPD_Reduction	✓	✓	✓
5	Main_Building_Type	✓	✓	✓
6	Orientation	✓	✓	✓
7	SHGC	✓	✓	✓
8	South_Wall_area(m^2)	✗	✗	✓
9	Wall_Area_(m^2)	✗	✓	✗
10	WWR_South	✓	✓	✓
11	WWR_North	✓	✓	✓
12	WWR_East	✓	✓	✓
13	WWR_West	✓	✓	✓

Using the optimize parameter operator, the implementation of different combinations of parameters' values on the modelling process for each tree criterion resulted in optimized models with acceptable performances which are shown in the below tables:

Table 14 - The optimized parameters based on three criterions

Operator	**Parameter**	**Parameter Value (optimized)**		
		gini_index	information_gain	gain_ratio
Random Forest	Maximal Depth	35	20	15
	Confidence	0.4	0.4	0.4
	Minimal Gain	0.01	0.01	0.01
	Minimal Leaf Size	3	3	3
	Prepruning Alternatives	7	10	5
	Minimal Size for Split	7	3	3
	Number of trees	70	60	70
Discretize	bins	9	13	24
Bagging	sample ratio	0.8	0.8	0.8
	iteration	10	10	10
Performance (Cross Validation)	Accuracy	**90.71%**	89.62%	90.37%

According to table 14 Gini index criterion results in a slightly better performance than the other two criterions.

Table 15 - The process performance with optimized RF parameters

Attribute Selection	Performance Type	Positive Class	Accuracy	Precision	Recall	F measure
Group 1	Cross Validation	1	90.71%	87.85%	75.84%	81.41%
		0		91.58%	96.16%	93.81%
Group 2	Cross Validation	1	92.93%	91.59%	81.08%	86.01%
		0		93.35%	97.27%	95.27%
Group 5	Cross Validation	1	96.09%	95.68%	89.56%	92.47%
		0		96.22%	98.52%	97.36%

Group 5 attributes includes the attributes which have the most influence on the EUI Classification in the RF model and are the outputs of Optimize Parameter operator in RapidMiner using Gini index as the criterion (table 13)

AUC and ROC are plotted for each attribute selection in the figures below:

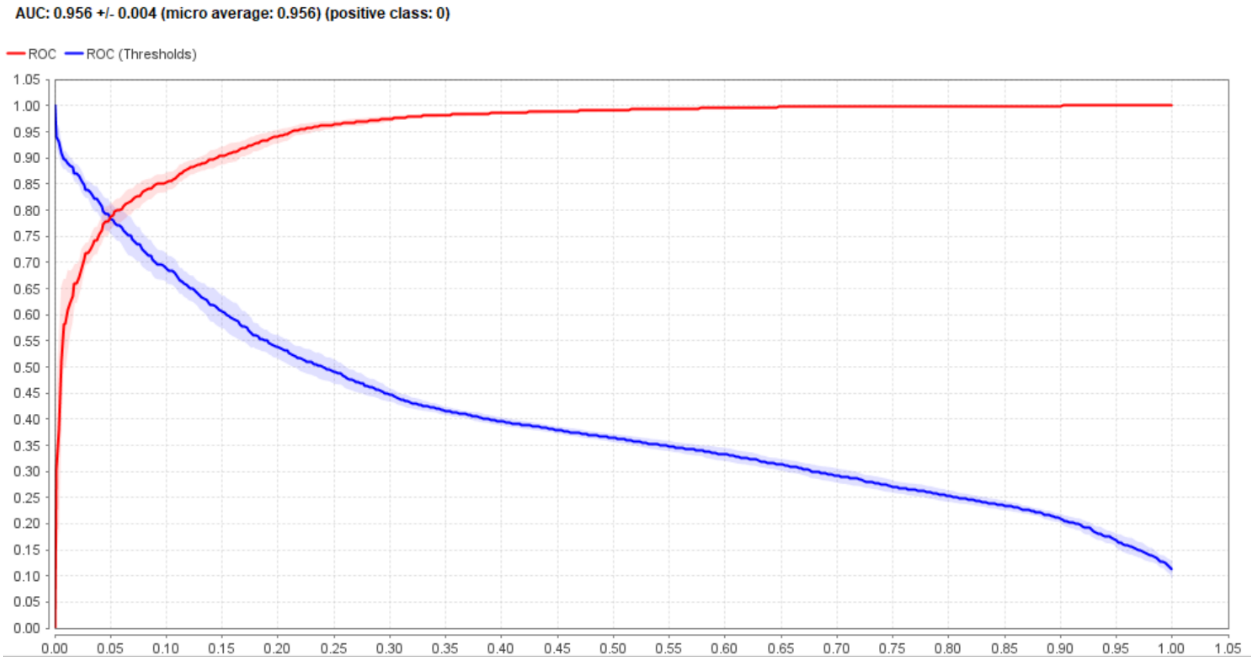

Figure 16 - AUC and ROC for the final optimized RF for EUI Classification (Group 1)

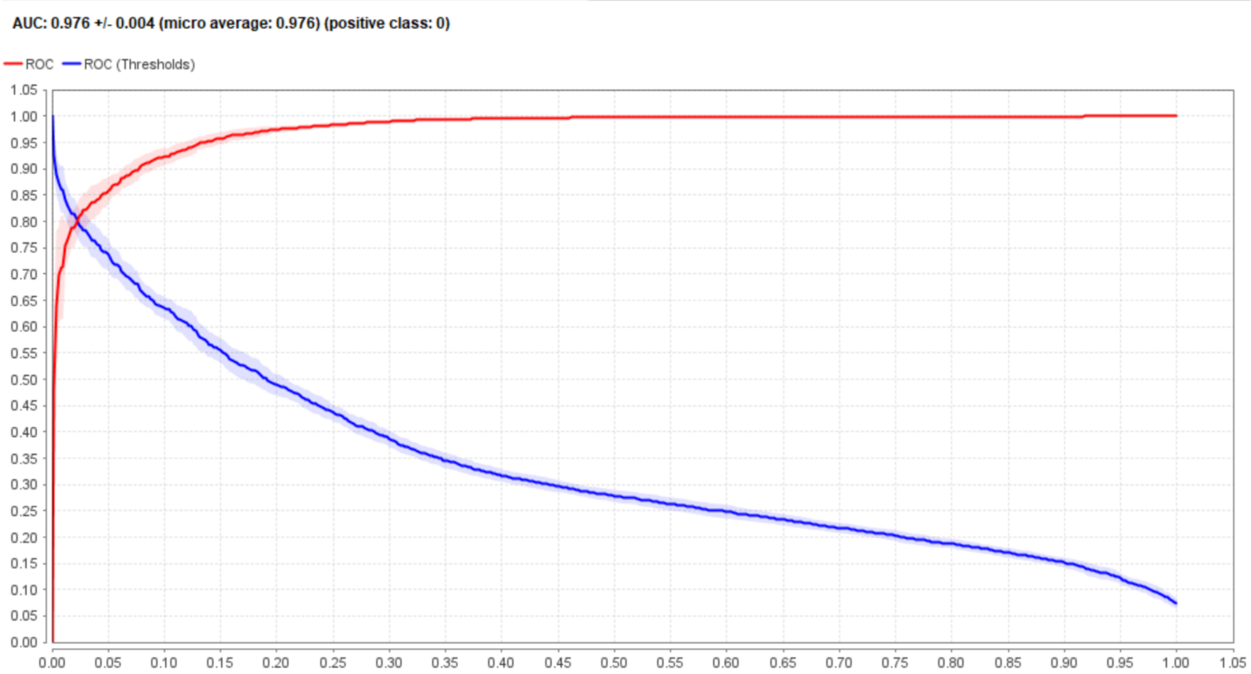

Figure 17 - AUC and ROC for the final optimized RF for EUI Classification (Group 2)

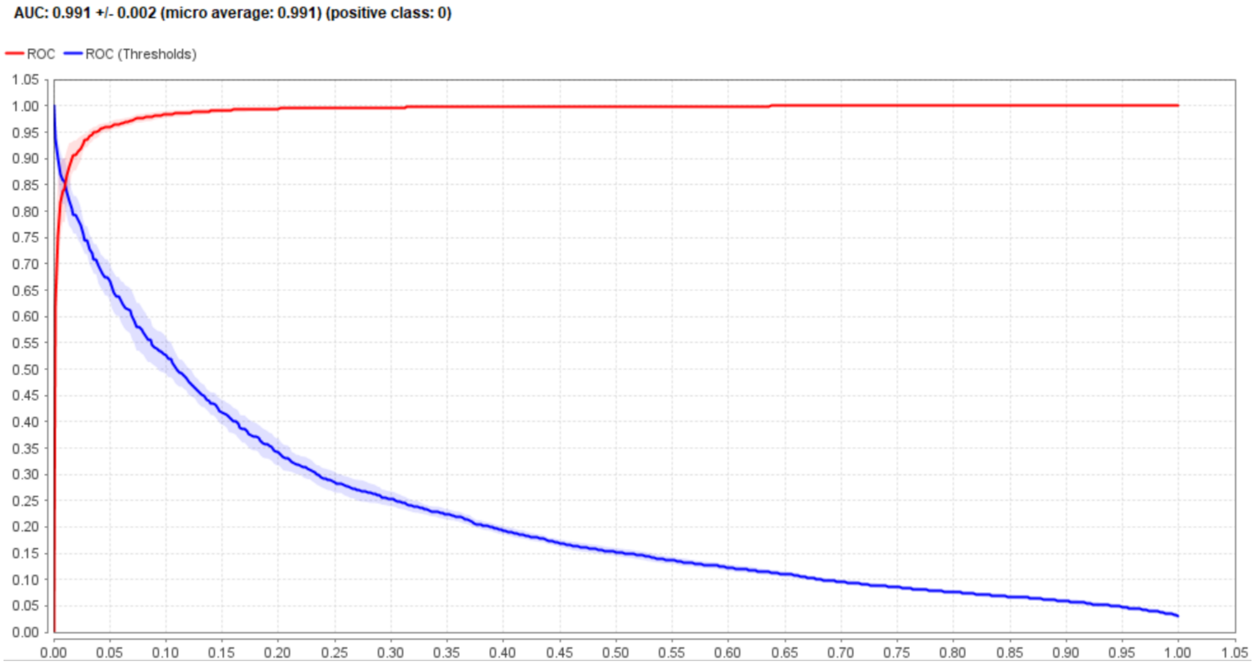

Figure 18 - AUC and ROC for the final optimized RF for EUI Classification (Group 5)

4.2 BCR

DT and RF models are trained, evaluated, and optimized separately.

4.2.1 Decision Tree

All the model parameters are selected the same as the EUI model which is already trained. One of the differences in BCR classification is the labeled attribute which is NWP Classification in this section. The other change in the implementation process is the attribute selection. As such four groups of attributes are selected to be able to compare the resulted models. These sets of attributes are listed in the following table:

Table 16 - Attribute selections for BCR Classification

Group 1	Group 2		Group 3	Group 6
All Attributes	East_Wall/Volume	South_Wall_%	HVAC_System	Heating_thermostat_Sch
	East_Wall_%	Stories	Lighting_Type	HVAC_System
	Height/Roof_area	Stories/Wall_area	LPD_Reduction	Lighting_Type
	Height/Wall_area	U-value	Main_Building_Type	LPD_Reduction
	Height_(m)	Volume/Roof_area	North_Wall/Volume	Main_Building_Type
	HVAC_System	Volume/Wall_area	Orientation	Orientation
	LPD_Reduction	Wall_area/Roof_area	R-value_Roof	SHGC
	North_Wall/Volume	West_Wall/Volume	SHGC	U-value_(W/(m2K))
	North_Wall_%	West_Wall_%	U-value_(W/(m2K))	WWR_East
	Orientation	WWR_East	Volume/Thermal_Zone	WWR_North
	R-value_Roof	WWR_North	WWR_East	WWR_South
	SHGC	WWR_South	WWR_North	WWR_West
	South_Wall/Volume	WWR_West	WWR_South	
			WWR_West	

In the above table, **Group** 1 refers to all the attributes generated by the data preparation and setting aside irrelevant attributes like ID which are only used to address each data point. **Group 2** attributes are those attributes which were supposed to be investigated through the project. **Group 3** includes attributes which have the most influence on determining the target values. **Group 6** includes the attributes which play a significant role in the classification of the target attribute (NPW Classification) which is used to address BCR Classification. NPW Classification is selected as the target attribute in BCR modeling.

Using the same DT parameters used for the EUI Classification model, the following results are generated:

Table 17 - BCR DT model performance

Attribute Selection	Performance Type	Positive Class	Accuracy	Precision	Recall	F measure
Group 1	Cross Validation	1	94.43%	95.99%	93.80%	94.88%
		0		92.62%	95.20%	93.90%
Group 2	Cross Validation	1	81.92%	75.69%	98.90%	85.75%
		0		97.85%	61.15%	75.27%
Group 3	Cross Validation	1	94.45%	95.96%	93.86%	94.90%
		0		92.69%	95.17%	93.91%
Group 6	Cross Validation	1	94.49%	95.99%	93.91%	94.94%
		0		92.75%	95.20%	93.96%

Figures 19 to 22 illustrates AUC and ROC metrics for each group which is mentioned in table 17:

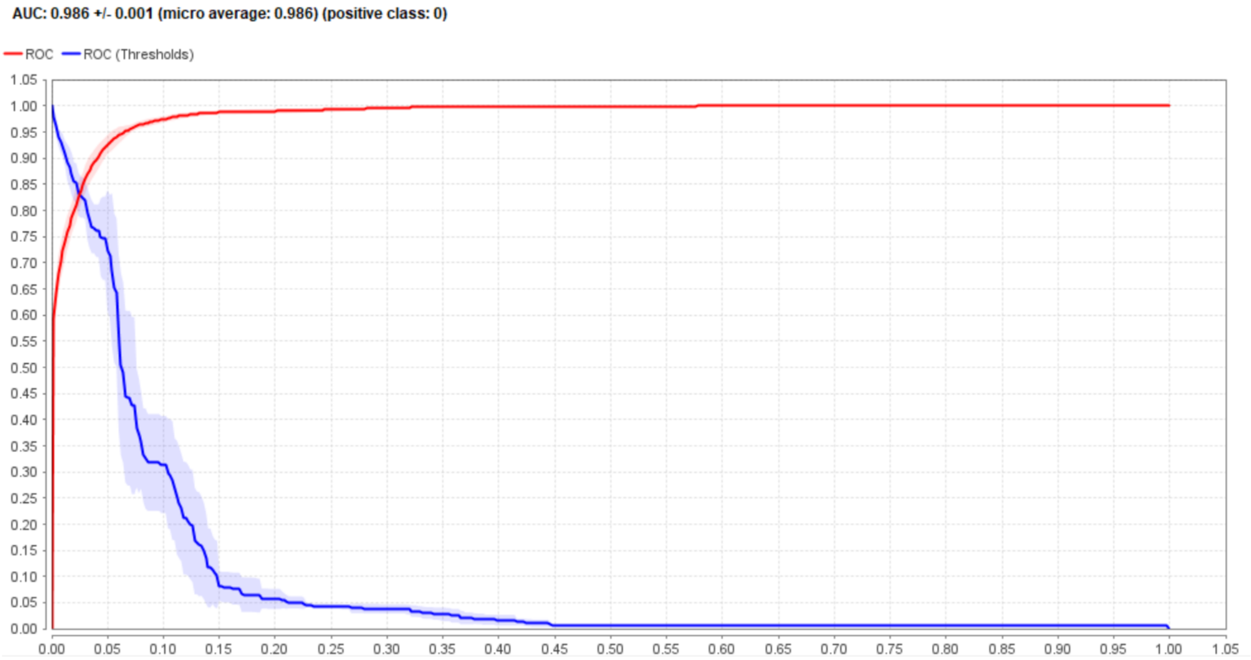

Figure 19 - AUC and ROC for the final optimized DT for BCR Classification (Group 1)

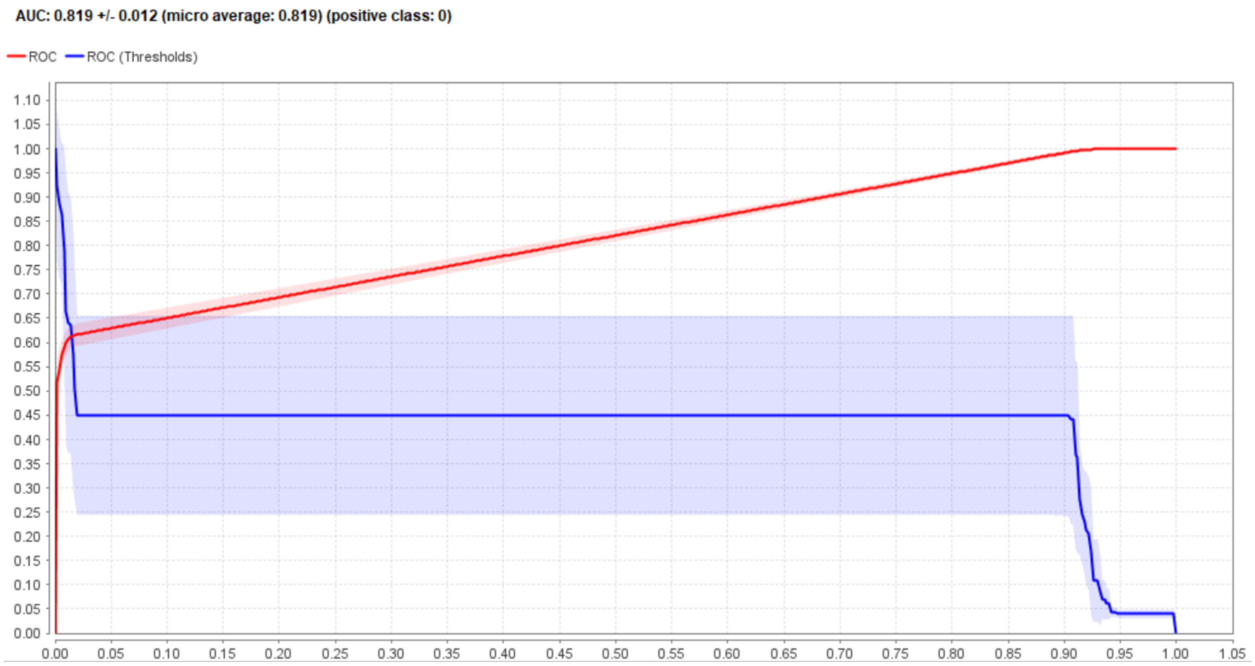

Figure 20 - AUC and ROC for the final optimized DT for BCR Classification (Group 2)

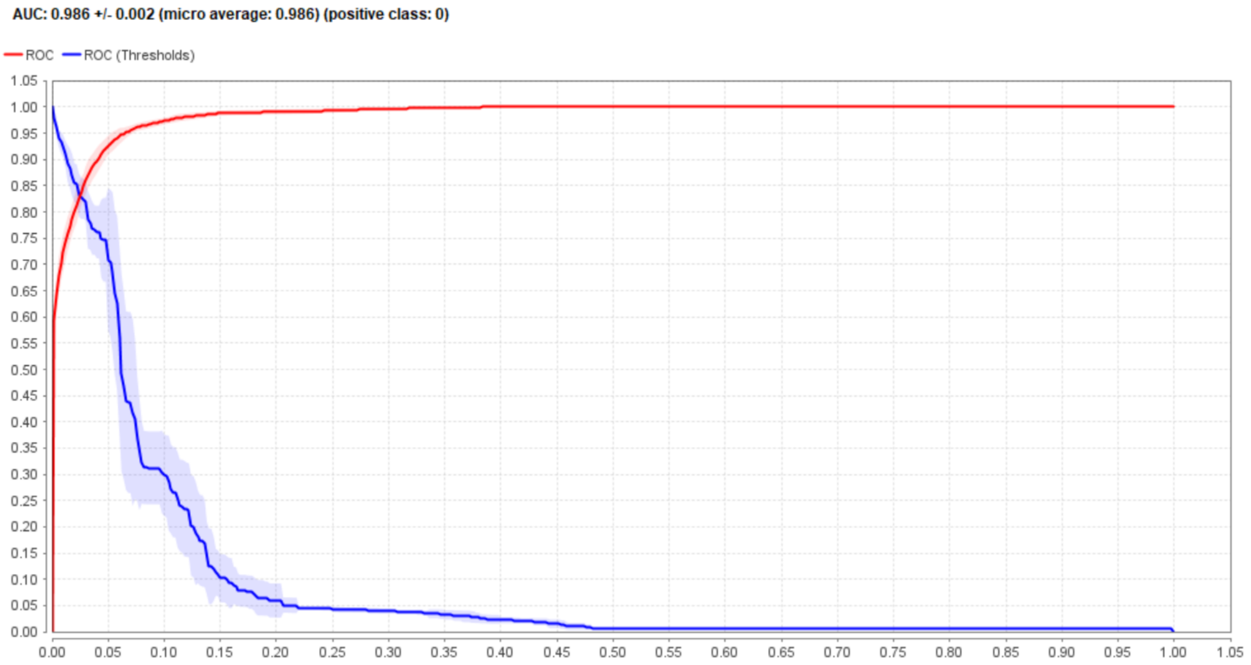

Figure 21 - AUC and ROC for the final optimized DT for BCR Classification (Group 3)

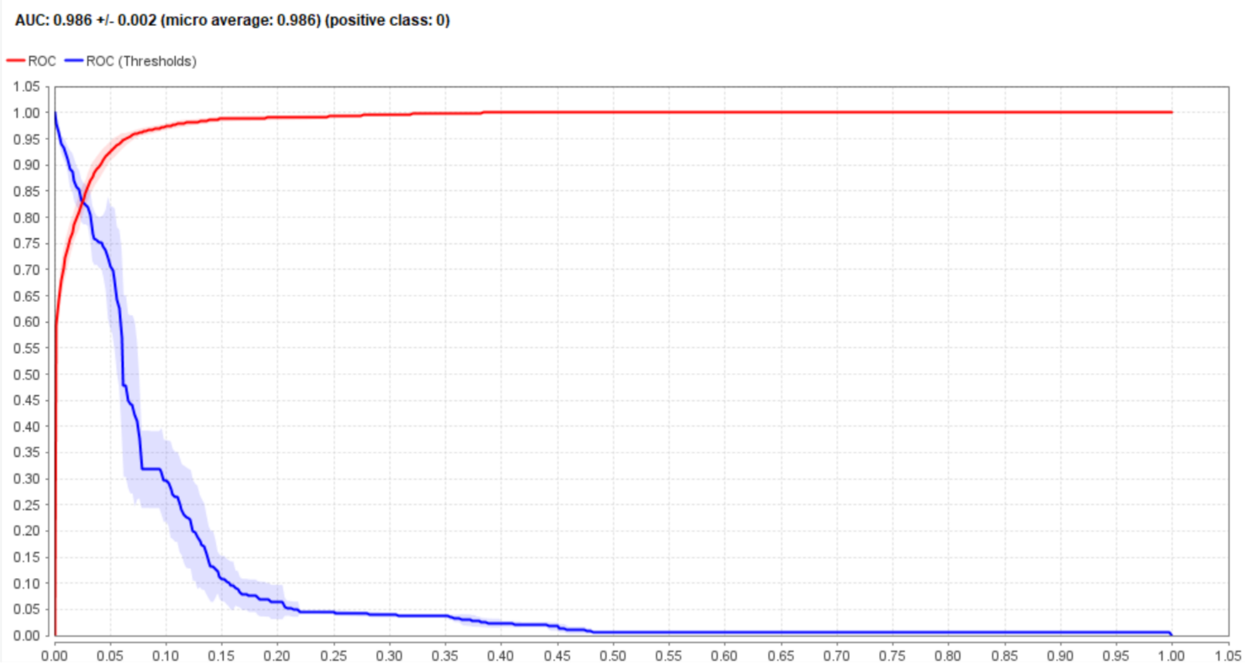

Figure 22 - AUC and ROC for the final optimized DT for BCR Classification (Group 6)

4.2.2 Random Forest

All the parameters of Random Forest model are the same as the EUI Classification model and similar to the DT modelling, there have been changes in target attribute as well as the attribute selection. Table 18 lists the sets of selected attributes.

Table 18 -Random Forest attribute selection for BCR classification

Group 1	Group 2	Group 5	Group 7
All Attributes	HVAC_System	Height_(m)	Heating %
	Lighting_Type	Lighting_Type	HVAC_System
	LPD_Reduction	LPD_Reduction	Lighting_Type
	Main_Building_Type	Main_Building_Type	LPD_Reduction
	Orientation	Orientation	Orientation
	R-value_Roof	SHGC	WWR_East
	SHGC	WWR_South	WWR_North
	South_Wall/Volume	WWR_North	WWR_South
	U-value_(W/(m2K))	WWR_East	WWR_West
	Volume/Wall_area	WWR_West	
	WWR_East		
	WWR_North		
	WWR_South		
	WWR_West		

In this table **Group 1** includes all attributes, **group 2** attributes are those attributes which were supposed to be investigated through the project, **Group 5** attributes are the determinative attributes for EUI classification, and **group 7** are the attributes which have the most influence on the target attribute (NPW Classification) classification. Using the cross-validation operator in RM, the following results are generated.

Table 19 - BCR RF model performance

Attribute Selection	**Performance Type**	**Positive Class**	**Accuracy**	**Precision**	**Recall**	**F measure**
Group 1	Cross Validation	1	90.78%	91.94%	91.24%	91.59%
		0		89.39%	90.22%	89.80%
Group 2	Cross Validation	1	92.74%	93.83%	92.92%	93.37%
		0		91.44%	92.52%	91.98%
Group 5	Cross Validation	1	96.23%	97.26%	95.84%	96.55%
		0		95.00%	96.70%	95.85%
Group 7	Cross Validation	1	96.53%	97.19%	96.47%	96.83%
		0		95.73%	96.60%	96.16%

For BCR Classification using RF, AUC and ROC graphs in figures 23 to 26 also confirm that group 5 and group 6 are better selection sets than group 1 and group 2 attributes:

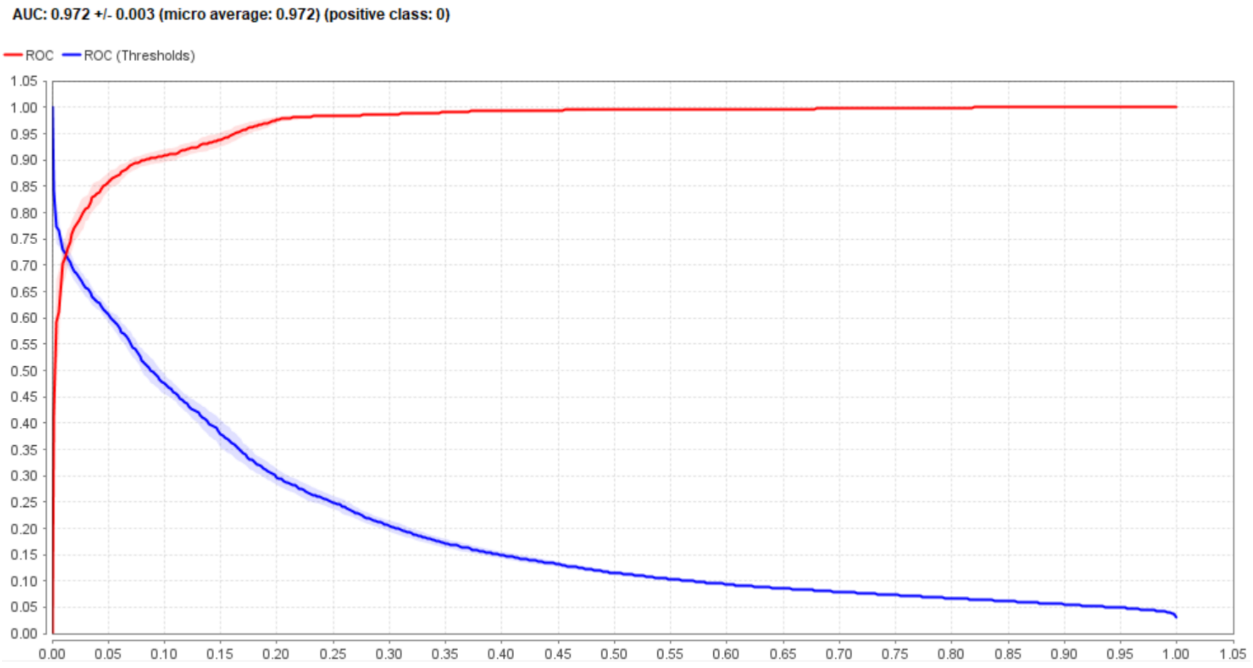

Figure 23 - AUC and ROC for the final optimized RF for BCR Classification (Group 1)

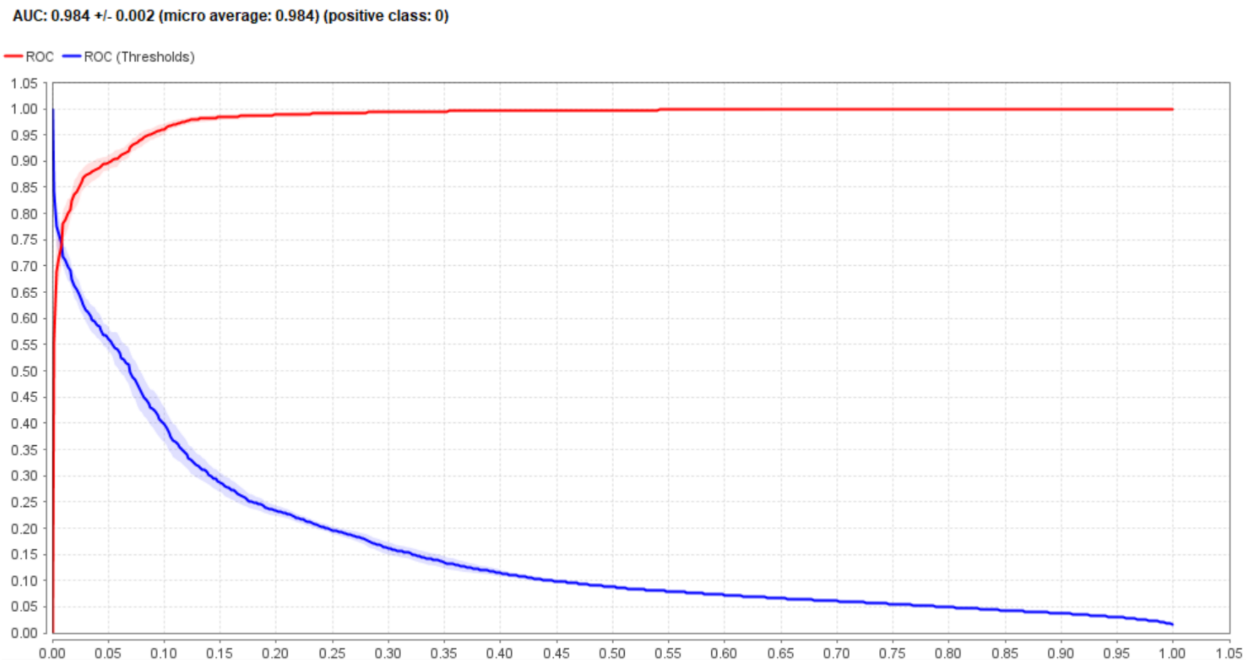

Figure 24 - AUC and ROC for the final optimized RF for BCR Classification (Group 2)

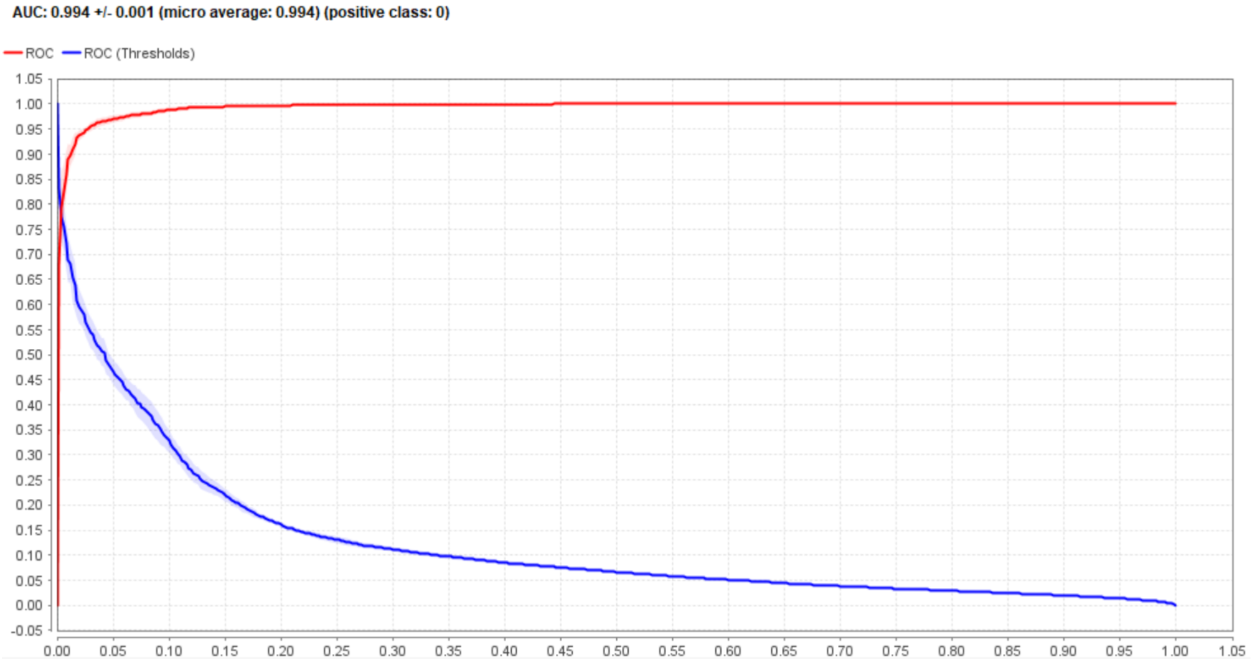

Figure 25 - AUC and ROC for the final optimized RF for BCR Classification (Group 5)

This document has been submitted by the student as a course project report, for evaluation.
It is NOT peer reviewed, and may NOT be cited as a scientific reference!

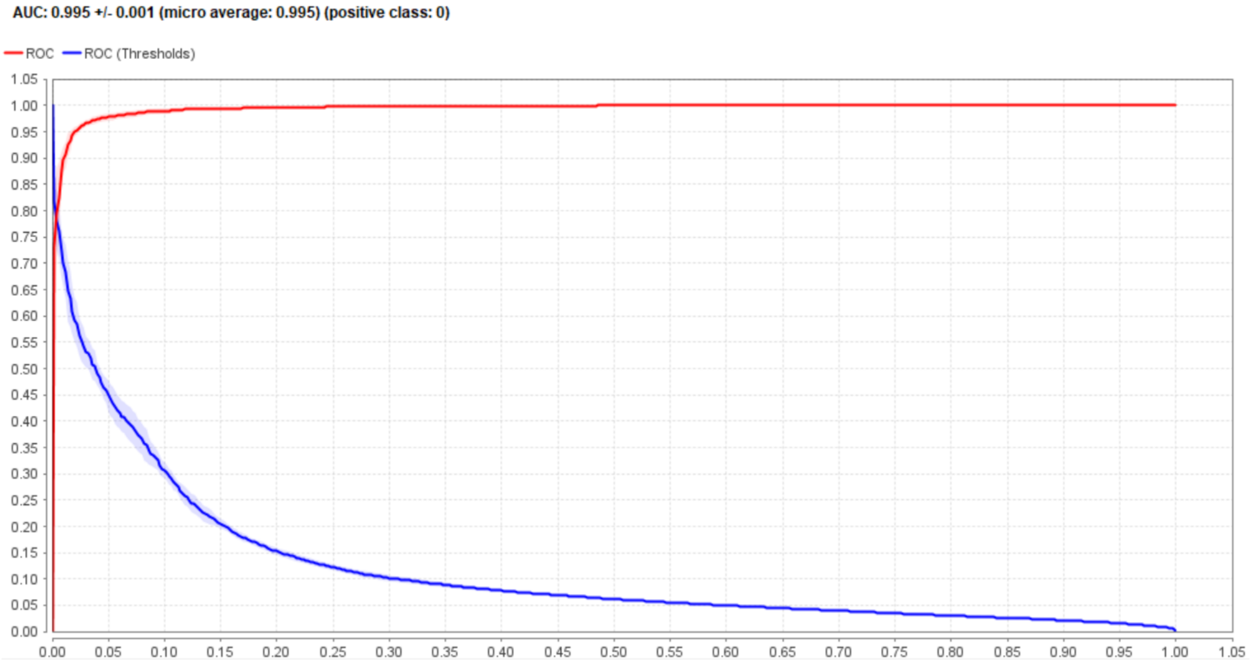

Figure 26 - AUC and ROC for the final optimized RF for BCR Classification (Group 7)

5. Summary and Discussion

The aim is to compare the best model results on DT and RF for both EUI and BCR classification and select the more reliable method among the two implemented ones which receives the design parameters of a building and confirms that EUI or BCR is either better or worse than a defined baseline which is the energy efficient or cost efficient design respectively. There are some measures to evaluate the performance of a model that helps to choose the best one which satisfies the objective. Three main metrics included in this study are accuracy, precision and recall.

Table 20 - Model performance results for the EUI

Method	**Attributes**	**Accuracy**	**class precision (positive class: 1)**	**class precision (positive class: 0)**	**class recall (positive class: 1)**	**class recall (positive class: 0)**	**AUC (positive class: 0)**	**F_measure (positive class: 0)**
Decision Tree	**Group 1**	**96.41%**	**94.99%**	**96.90%**	**91.41%**	**98.23%**	**99.10%**	**97.56%**
	Group 2	90.48%	76.65%	97.11%	92.71%	89.66%	97.20%	93.23%
	Group 3	**96.47%**	**95.05%**	**96.97%**	**91.62%**	**98.25%**	**99.20%**	**97.61%**
Random Forest	Group 1	90.71%	87.85%	91.58%	75.84%	96.16%	95.60%	93.81%
	Group 2	92.93%	91.59%	93.35%	81.08%	97.27%	97.60%	95.27%
	Group 5	**96.09%**	**95.68%**	**96.22%**	**89.56%**	**98.52%**	**99.10%**	**97.36%**

Considering accuracy as the criteria for the best model among the implemented ones, DT has the highest figure with 96.47% while selecting Group 3 attributes. Accuracy shows the portion of correct predictions of both classes without the ability to distinguish the performance in each class as the aim of this study is to determine the class representing EUI better than the baseline (positive class=0).

Considering precision (positive class = 0=EUI better than the baseline) as the criteria for the best model among the implemented ones, DT has the highest figure on Group 2 attributes with 97.11%.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Precision shows the portion of relevant cases, meaning that what portion of the predicted EUI better than the baseline examples, was in fact, better than the baseline. Precision does not consider the portion which was better than the baseline (i.e. true positives + false negatives) but did not predict in the model, so it is not enough to choose the best model based on this metric. There is another measure called recall, which shows what portion of relevant cases is found, i.e. what percentage of actual EUI better than the baseline examples were classified correctly.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Considering recall (positive class = 0=EUI better than the baseline) as the criteria for the best model among the implemented ones, RF model on Group 6 attributes has the highest figure with 97.52%.

In conclusion, for EUI Classification, among DT and RF, RF provides a model with slightly better performance.

Table 21 - Model performance results for BCR

Method	Attributes	Accuracy	precision (positive class: 1)	precision (positive class: 0)	recall (positive class: 1)	recall (positive class: 0)	AUC (positive class: 0)	F_measure (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
Decision Tree	Group 1	94.43%	95.99%	92.62%	93.80%	95.20%	98.60%	93.90%	95.20%	93.80%
	Group 2	81.92%	75.69%	97.85%	98.90%	61.15%	61.15%	75.27%	61.15%	98.90%
	Group 3	94.45%	95.96%	92.69%	93.86%	95.17%	98.60%	93.91%	95.17%	93.86%
	Group 6	**94.49%**	**95.99%**	**92.75%**	**93.91%**	**95.20%**	**98.60%**	**93.96%**	**95.20%**	**93.91%**
Random Forest	Group 1	90.78%	91.94%	89.39%	91.24%	90.22%	97.20%	89.80%	90.22%	91.24%
	Group 2	92.74%	93.83%	91.44%	92.92%	92.52%	98.40%	91.98%	92.52%	92.92%
	Group 5	**96.23%**	**97.26%**	**95.00%**	**95.84%**	**96.70%**	**96.20%**	**95.85%**	**96.44%**	**95.88%**
	Group 7	**96.53%**	**97.19%**	**95.73%**	**96.47%**	**96.60%**	**99.50%**	**96.16%**	**96.60%**	**96.47%**

For BCR Classification, considering accuracy as the deciding criteria, RF on group 7 has the highest accuracy of 96.53%. Since in BCR, the recall for 1 as the positive class is more important as it represents better than the baseline BCR, RF on group 7 is considered as the better performing model.

Referring to table 20 and table 21, it can be observed that RF on group 5 has reliable results for both EUI and BCR classifications and it generate an interesting results showing that by having the values of group 5 attributes which consists of only ten attributes, using the optimized RF in this study, reliable classifications for EUI and BCR can be achieved. These ten attributes are:

- Height_(m)
- Lighting_Type
- LPD_Reduction

- Main_Building_Type
- Orientation
- SHGC
- WWR_South
- WWR_North
- WWR_East
- WWR_West

5.1 Future Works:

The dataset for training the models could be more comprehensive covering a wider range of each attributes as well as identifying more probable influencing factors.

There are more complex modeling methods such as KNN, ANN, Naïve Bayes, and etc. that might also create reliable results. Moreover K-Means Clustering could be applied on the dataset for data preparation.

The dataset used in this project was the result of multiple simulation scenarios and it would be interesting to train and test the models using the real-world datasets.

6. References:

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings*, vol. 40, no. 3, pp. 394–398, 2008, doi: https://doi.org/10.1016/j.enbuild.2007.03.007.

[2] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy and Buildings*, vol. 42, no. 10, pp. 1637–1646, 2010, doi: https://doi.org/10.1016/j.enbuild.2010.04.006.

[3] B. Yildiz, J. Bilbao, and A. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017, doi: 10.1016/j.rser.2017.02.023.

[4] H. Kim, A. Stumpf, and W. Kim, "Analysis of an energy efficient building design through data mining approach," *Automation in Construction*, vol. 20, no. 1, pp. 37–43, 2011, doi: https://doi.org/10.1016/j.autcon.2010.07.006.

[5] F. Xiao and C. Fan, "Data mining in building automation system for improving building operational performance," *Energy and Buildings*, vol. 75, pp. 109–118, 2014, doi: https://doi.org/10.1016/j.enbuild.2014.02.005.

[6] A. Capozzoli, D. Grassi, M. S. Piscitelli, and G. Serale, "Discovering Knowledge from a Residential Building Stock through Data Mining Analysis for Engineering Sustainability," *Energy Procedia*, vol. 83, pp. 370–379, 2015, doi: https://doi.org/10.1016/j.egypro.2015.12.212.

[7] A. Ashari, I. Paryudi, and A. M. Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *International Journal of Advanced Computer Science and Applications*, vol. 4, 2013, doi: 10.14569/IJACSA.2013.041105.

[8] C. C. Yang, C. S. Soh, and V. V. Yap, "A systematic approach in appliance disaggregation using k-nearest neighbours and naive Bayes classifiers for energy efficiency," *Energy Efficiency*, vol. 11, no. 1, pp. 239–259, Jan. 2018, doi: 10.1007/s12053-017-9561-0.

[9] *RapidMiner*. Rapidminer.