

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It is NOT peer reviewed, and may NOT be cited as a scientific reference!



Department of Building,
Civil, & Environmental Engineering

Building Baseline Classifier

(In terms of Energy Usage Intensity (EUI) and Benefit Cost Ratio (BCR))

Mohammadjavad Anbia

40079574

ENGR 6991 Project and Report III

Supervisor: Dr. Mazdak Nik-bakht

Jan - Aug 2020

Abstract

This project aims to increase the energy usage efficiency of buildings by discovering and classifying the effective energy-related factors for different design alternatives of buildings. The design alternatives are obtained from several softwares such as Building Information Modelling (BIM) related software like Autodesk Revit and open-source cloud computing technologies like Open Studio to evaluate the energy usage and cost analysis of the buildings.

This report focuses on the classification of both EUI and BCR data by using four different classification methods; K-nearest Neighbors (KNN), Naïve Bayes (NB), Support vector machine (SVM) and Artificial Neural Network (ANN).

Each of these two target attributes, EUI and BCR, are divided into two groups of better or worse than the previously defined baselines, which result in classification models that can predict whether a building is better or worse than the EUI or BCR baselines.

The primary software for this classification is “RapidMiner” machine learning software. Also, the CRISP-DM machine learning structure is used to plan and shape the procedure of the project.

In the end, the performance of each model is evaluated and compared with each other to achieve a model with the highest performance level in terms of proximity metrics; accuracy, recall, precision and AUC (Area Under Curve). The result of the comparison concluded that the ANN has the highest performance between the models even though the other models resulted in acceptable performances.

Table of Contents

List of Figures	V
List of Tables	VI
Chapter 1.....	1
1. Introduction.....	1
1.1 Motivation and Background	1
1.2 Problem Statement and Objectives	2
Chapter 2.....	3
2. Literature Review.....	3
2.1 A review and analysis of regression and machine learning models on	3
commercial building electricity load forecasting[2]	3
2.2 Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in	4
Searching Alternative Design in an Energy Simulation Tool[3]	4
2.3 A systematic approach in appliance disaggregation using k-nearest neighbours and naive Bayes	4
classifiers for energy efficiency[4]	4
Chapter 3.....	5
3. Dataset.....	5
Chapter 4.....	7
4. Methodology	7
4.1 Artificial Neural Network (ANN).....	7
4.1.1 Data preparation.....	8
4.1.2 Modelling.....	9
4.1.3 Evaluation.....	9
4.2 K-Nearest Neighbor and Naïve Bayes	9
4.2.1 Data preparation and Pre-processing	10
4.2.2 Modelling:	13
4.2.3 Evaluation.....	14
4.3 Support Vector Machine	15
4.3.1 Data preparation and Pre-processing	15
4.3.2 SVM Modelling.....	15
4.3.3 Evaluation	16
Chapter 5.....	17
5. Implementation and Results.....	17

5.1 Artificial Neural Network	17
5.2 K-Nearest Neighbor.....	19
5.3 Naïve Bayes.....	20
5.3.1 First Set of Selected Attributes	20
5.3.2 Second Set of Selected Attributes	22
5.4 Support Vector Machine	23
Chapter 6.....	26
6. Discussion	26
6.1 Energy Use Intensity	26
6.2 Benefit to Cost Ratio	27
Chapter 7.....	29
7. Conclusion	29
8. References.....	30

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!

List of Figures

Figure 1 A general ANN with two hidden layers and its main components [5].....	7
Figure 2 KNN/SVM/ANN implementation process	8
Figure 3 NB implementation process.....	9
Figure 4 KNN performance evaluation in the Rapidminer	14
Figure 5 Neural Network area under curve and receiver operating characteristic (ROC) for EUI classification.....	18
Figure 6 Neural Network area under curve and receiver operating characteristic (ROC) for EUI classification.....	18
Figure 7 KNN area under curve and receiver operating characteristic (ROC) for EUI classification.....	19
Figure 8 KNN area under curve and receiver operating characteristic (ROC) for BCR classification.....	20
Figure 9 NB area under curve and receiver operating characteristic (ROC) for EUI classification - First set of attributes	21
Figure 10 NB area under curve and receiver operating characteristic (ROC) for EUI classification - Second set of attributes	21
Figure 11 NB area under curve and receiver operating characteristic (ROC) for BCR classification - First set of attributes	22
Figure 12 NB area under curve and receiver operating characteristic (ROC) for BCR classification - Second set of attributes	23
Figure 13 SVM area under curve and receiver operating characteristic (ROC) for EUI classification.....	25
Figure 14 SVM area under curve and receiver operating characteristic (ROC) for BCR classification	25
Figure 15 EUI Classification models Performance	27
Figure 16 BCR Classification models Performance	28

List of Tables

Table 1 List of all attributes	5
Table 2 Filtered Attributes	10
Table 3 First set of selection for NB	12
Table 4 Second set of selection for NB (efficient attributes).....	13
Table 5 Sorted amounts of k by Recall using Oprimize operator.....	14
Table 6 Neural Network performance results for EUI classification	17
Table 7 Neural Network performance results for NPW classification	17
Table 8 KNN performance results for EUI classification.....	19
Table 9 KNN performance results for NPW classification	19
Table 10 NB performance results for EUI classification.....	20
Table 11 NB performance results for NPW classification.....	20
Table 12 NB performance results for EUI classification.....	22
Table 13 NB performance results for NPW classification.....	22
Table 14 SVM most effective attributes sorted weight table.....	23
Table 15 SVM performance results for EUI classification.....	24
Table 16 SVM performance results for NPW classification	24
Table 17: Modelling Performances Results for EUI Models.....	26
Table 18 Modelling Performances Results for BCR Models	27

Chapter 1

1. Introduction

1.1 Motivation and Background

Usually, the energy simulation is done during the late phases of building design when most of the fundamental element's designs are finished. The energy simulation in the early stages of design helps to have more energy-related aspects included in the design and also to have a more effective impact on the life cycle manner of buildings. This project is a continuation of "*Economy-energy trade-off automation – A decision support system for building design development*"[1] article, which focuses on the integration of the energy simulation and cost analysis at the early schematic stage of the design.

Many design alternatives are created to perform the simulation, by using several software such as Building Information Modelling (BIM) related software like Autodesk Revit and open-source cloud computing technologies like Open Studio to evaluate the energy usage and cost analysis of the buildings. In order to compare the different designs in terms of energy usage and the economic analysis, a baseline is assumed based on the minimum acceptable limit in codes and standards for every examined attribute.

The classification of building-related parameters such as architectural specifications, lightning power, building usage and energy efficiency parameters like R-value and U-value could help to distinguish the life cycle efficient designs with the inefficient ones. The classification is performed based on two main target criteria; Energy Usage Intensity (EUI) and BCR (Benefit to Cost Ratio). Then these two target attributes are divided into two groups of better or worse than the previously defined baselines, which result in classification models that can predict whether a building is better or worse than the EUI or BCR baselines.

The classification process in this project is performed for both EUI and BCR data by using four different classification methods; K-nearest Neighbors (KNN), Naïve Bayes (NB), Support vector machine (SVM) and Artificial Neural Network (ANN). The primary software for this classification is "RapidMiner" machine learning software. Also, the CRISP-DM machine learning structure is used to plan and shape the procedure of the project.

1.2 Problem Statement and Objectives

This project aims to classify the building-related attributes in terms of EUI and BCR better or worse than the specified baselines. This classification could help the designers to have an executive plan for developing efficient energy-related elements to reduce the energy usage and life cycle cost of buildings from the very early stages of designs. In this regard, four different classification models of KNN, NB, SVM and ANN data mining techniques are used, which could also predict the EUI and BCR better or worse the specified baseline for an unknown building design.

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!

Chapter 2

2. Literature Review

In the literature, some articles that classify the building effective specifications that use the afore-mentioned classification methods, which are KNN, NB, SVM and ANN, are gathered for better understanding the purpose of this project. It should be mentioned that this project is a part of a primary project which uses decision tree and random forest methods as well. However, in this report, only the four mentioned methods are discussed.

2.1 A review and analysis of regression and machine learning models on commercial building electricity load forecasting[2]

B. Yildiz, J.I. Bilbao and A.B. Sproul reviewed different regression and machine learning models that forecast the electricity load of commercial buildings and also the regression variables, which could improve the accuracy of their performance. The regression models in the study are Single and Multilinear regression, regression trees (RT) and Support Vector Regression (SVR). The machine learning models are dynamic ANN models, which were Levenberg-Marquardt Backpropagation (LM) and the Bayesian Regularization Backpropagation (BR). Then, the different models were compared in terms of their next day's hourly electricity load for different seasons of the year. Four main performance parameters of Root Mean Square Error (RMSR), Mean Absolute Percentage Error (MAPE), R^2 and Mean Bias Error (MBE) are used for the comparison.

Most of the machine learning models showed a better forecast performance and accuracy than the multilinear regression models even though the regression models are more comfortable and straightforward to implement. Between the machine learning models, the ANN models of LM and BR resulted in higher performance with RMSE= 1.92 and 1.87, MAPE= 1.36 and 1.33, R^2 = 0.99 for both and MBE (%) = 0.03 and -0.01, respectively. Besides, some factors found out to influence the accuracy of the models, such as increasing the number of examples in the training sets and eliminating the irrelevant attribute parameters.

2.2 Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool[3]

Ahmad Ashari, Iman Paryudi, A Min Tjoa classified different building designs' various specifications such as wall area and floor area by using Weka data mining tool. The classifiers in the study are Naïve Bayes, Decision Tree, and k-Nearest Neighbor methods. The fastest method is a decision tree rather than the other two since KNN and NB methods need calculations, and it slows down the process of classifying. The measuring of the model's performance is done based on Precision, Recall, F-Measure, Accuracy, and AUC. In the end, the NB model concluded a higher accuracy of 0.737 rather than the DT and KNN with accuracies of 0.589 and 0.567, respectively. The other performance indices of NB models are; Precision=0.799, Recall=0.794, F-Measure=0.78, and AUC=0.605. The high accuracy of the NB model is because the dependencies of attributes distributed over classes rather than the attributes themselves.

2.3 A systematic approach in appliance disaggregation using k-nearest neighbours and naive Bayes classifiers for energy efficiency[4]

Chuan Choong Yang & Chit Siang Soh & Vooi Voon Yap aim to improve energy efficiency by using the appliances energy usage feedbacks and behavioural consumption changes. They used two data mining methods of k-nearest neighbours (k-NN) and Naive Bayes (NB) that are adopted for a REDD public dataset for classification. By evaluating the performances of the models, it is concluded that the KNN models concluded higher performances than NB models with the accuracies of higher than 78% and precision and recall of more than 0.3 and 0.4, respectively.

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!

Chapter 3

3. Dataset

The dataset for data mining includes 57 attributes and 14687 data points. Table 1 is the list of attributes with their spread and type.

Table 1 List of all attributes

Main Attributes	Type	Spread				Category
		Min, Least	Max, Most	Mean	Deviation	
Data_point_ID	Integer	1	14688	-	-	ID
Building_Model	Polynomial	17 types		-	-	General Information
Multiuse_Building? (1: yes; 0: no)	Binominal	0	0	-	-	
Main_Building_Type	Polynomial	12 types		-	-	
Rise_(low_or_high)	Binominal	high(5184)	low(9504)	-	-	
Stories	Integer	1	13	3.47	3.41	
Data_center_in_the_Building? (1: yes; 0: no)	Binominal	1(864)	0(13824)	-	-	Architectural parameters
Height_(m)	Real	3.05	51.48	12.85	12.33	
Building_Volume_(m^3)	Real	708.37	126016.35	31204.6	35682.6	
Roof_footprint_(ft^2)	Integer	2786	126672	30773.12	30577.1	
Roof_footprint_(m^2)	Real	258.83	11768.21	2858.9	2840.7	
Orientation	Polynomial	90(1574)	0(7344)	-	-	
Wall_Area_(ft^2)	Real	1721	74849	23939	21734.3	
Wall_Area_(m^2)	Real	159.89	6953.7	2224	2019.2	
South_Wall_area(m^2)	Real	46.47	3476.86	976.38	973.73	
North_Wall_area(m^2)	Real	46.47	3476.86	972.13	963.93	
East_Wall_area(m^2)	Real	46.47	2317.91	513.48	590.54	
West_Wall_area(m^2)	Real	46.47	2317.91	518.45	600.12	
South_Wall_%	Real	16.1776	57.3789	40.98	11.36	
North_Wall_%	Real	16.1776	56.8327	40.85	11.02	
East_Wall_%	Real	7.6564	35.5911	22.88	7.37	
West_Wall_%	Real	7.6564	35.5911	22.96	7.4	
WWR_South	Real	0.0019	0.991	0.3092	0.1928	
WWR_North	Real	0	0.991	0.2432	0.2351	
WWR_East	Real	0	0.991	0.2561	0.2273	
WWR_West	Real	0	0.991	0.2554	0.2268	
Stories/Roof_area	Real	0.0001	0.0128	0.0025	0.003	
Stories/Wall_area	Real	0.0004	0.0063	0.0022	0.0016	
Height/Roof_area	Real	0.0006	0.0389	0.0083	0.0089	
Height/Wall_area	Real	0.002	0.0191	0.0077	0.0045	
Volume/Roof_area	Real	2.6038	35.367	11.15	9.36	
Volume/Wall_area	Real	4.4305	24.5455	11.79	5.15	
Volume/Thermal_Zone	Real	98.3239	13747.9052	2123.13	3424.28	
Wall_area/Roof_area	Real	0.2376	3.4513	1.04	0.8651	
South_Wall/Volume	Real	0.0099	0.0747	0.041	0.0179	
North_Wall/Volume	Real	0.0099	0.0701	0.0408	0.0174	
East_Wall/Volume	Real	0.0087	0.0656	0.0231	0.014	
West_Wall/Volume	Real	0.0087	0.0656	0.0232	0.014	

Main Attributes	Type	Spread				Category
		Min, Least	Max, Most	Mean	Deviation	
U-value_(W/(m2K))	Real	1.624	3.122	2.49	0.4949	Energy efficiency
R-value_Roof	Integer	21	65	46.19	12.68	
R-value_Wall	Integer	16	41	28.32	7.27	
Heating %	Integer	24	64	47.65	12.74	
SHGC	Real	0.476	0.762	0.6439	0.0958	
Number_of_glazings(0: dbl; 1: trp)	Binominal	6 types of Double(6768),		-	-	Window type
Air_or_Arg_(0: air; 1: arg)	Binominal	1(4752)	0(9936)	-	-	
Lighting_Type	Polynomial	456), High_Bay_Low_Bay (3-		-	-	Lighting power
LPD_Reduction	Integer	1	37	20.1	10.39	
Thermal_Zones	Integer	2	118	34.2	33.2	HVAC system design
HVAC_System	Polynomial	11 types of HVAC Systems		-	-	
Heating_thermostat_Sch(hr/week)	Integer	72	168	125.35	36.53	
Cooling_thermostat_Sch(hr/week)	Integer	60	168	135.76	40.1	
Heating_Setpoint_Temp	Integer	15.6	21.7	20.8	1.32	
Cooling_Setpoint_Temp	Integer	22.2	29.44	24.23	1.38	
EUI	Real	11.8	213.5	67.18	40.72	Target Attributes
EUI Classification (1: Better than the baseline; 0: Worse than the baseline)	Binominal	0(2267)	1(7525)	-	-	
NPW	Real	184171.7	25994729.92	3854018	5253444	
NPW Classification (1: Better than the baseline, 0: worse than the baseline)	Binominal	0 (6608)	1 (8080)			

The target attributes of the project are EUI (Energy Usage Intensity) and BCR (Benefit Cost Ratio). Also, we divided the EUI and BCR values into binominal formats to compare the different designs of buildings to the baselines.

The dataset shows the different factors which could have effects on the target attributes of different designs in different building categories. The purpose will be to classify the EUI and BCR (as the target values) based on various design factors. In addition, it could be found which attributes have more influences on EUI and BCR in terms of being better or worse than the specified baselines.

Chapter 4

4. Methodology

The methodology of this report is divided by the four beforementioned methods, which are explained entirely below.

4.1 Artificial Neural Network (ANN)

An artificial neural network is a numerical model that is inspired by the way the human brain's neural network processes information. An ANN consists of an interconnected group of artificial neurons (Abhinav Saxena, 2007). ANN method analyses the data using interrelated networks of simple or uniform components. In general, it is a kind of network that is adjusted according to the information fed as the input to the network during the learning procedure. Modern neural networks are usually applied for sophisticated relations of the information or to reach patterns in data.[5]

On the one hand, a feed-forward neural network is a network where connections between the elements do not make a cycle or a loop. In this system, the information only flows forward. (RapidMiner, n.d.) On the other hand, a backpropagation needs to be performed to decrease the number of biases and adjust the weights in predicted values.[6]

Figure 1 describes a general ANN and a short description of its components.

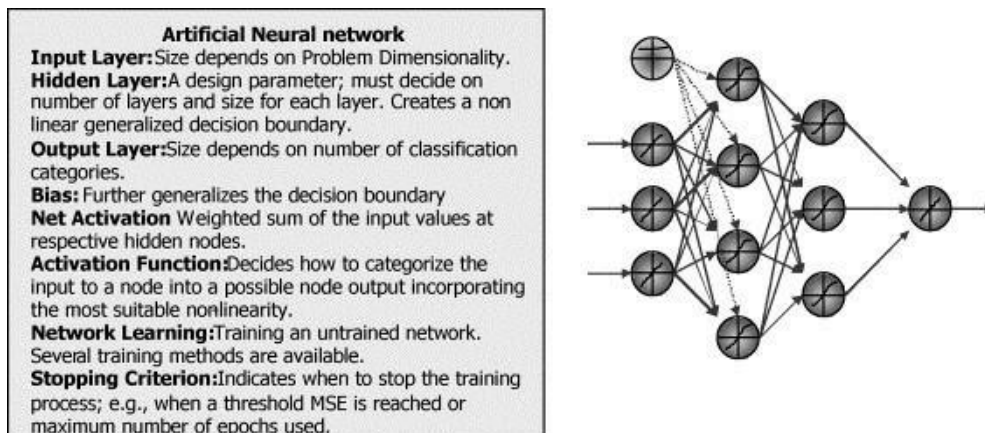


Figure 1 A general ANN with two hidden layers and its main components [5]

The CRISP-DM, which is a structured approach for planning a data mining project, is chosen for implementing the project. The flowchart below shows the different steps of implementation.

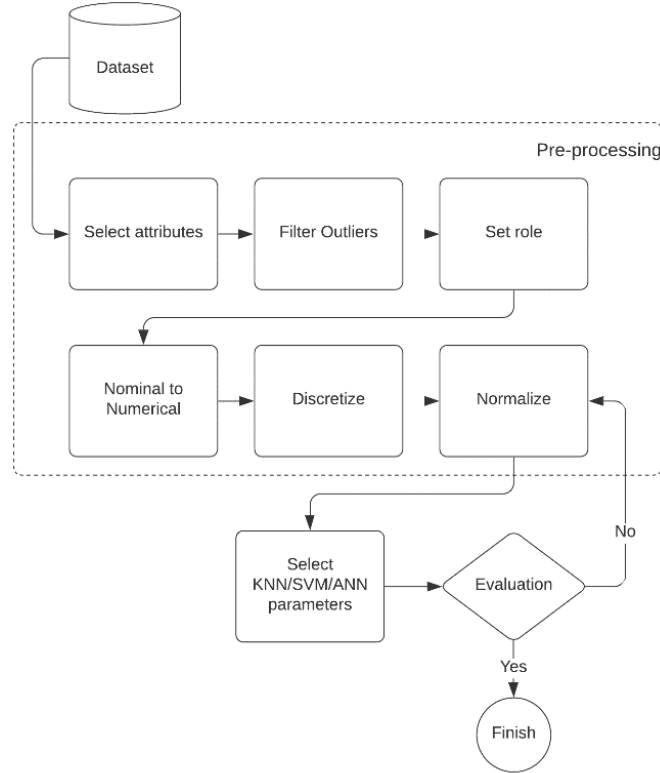


Figure 2 KNN/SVM/ANN implementation process

4.1.1 Data preparation

The following steps need to be performed for data preparation, as can be seen from the pre-processing part of figure 1.

4.1.1.1. Missing/Inconsistent values

Like all CRISP-DM procedures, the Missing/Inconsistent values and the outliers are identified and dealt with.

4.1.1.2. Nominal to Numerical

As the ANN model is a mathematical model, it is important to make sure that all the attributes are in a numerical format, so all the nominal attributes should turn into a numerical format while paying attention to whether or not the categorical values in each attribute have an order. For those values that are without any order, dummy coding should be used.

4.1.1.3. Set Role

Since the neural network operator is being used for classification, the target attribute should be labelled using the “Set Role” operator.

4.1.1.4. Normalize

As the activation function in ANN uses the sigmoid probability function, it is needed for all the attributes values to be on the same scale (-1 to +1); therefore, they are normalized using the “Normalize” operator.

4.1.2 Modelling

The Neural Network operator uses a forward propagation trend for performing the mathematical ANN modelling. In this numerical model, some random values will be selected for the input attributes of the dataset and send them to a hidden layer to calculate the output predicted values by an activation function based on the assumed weight and resulted in biases in the calculations.

The next phase is iteration, as this process will be continued to the results of actual values from the training set get closer to the predicted values of the output.

4.1.3 Evaluation

Cross-validation is used in addition to a binominal performance to evaluate the models as well as to avoid overfitting.

4.2 K-Nearest Neighbor and Naïve Bayes

The second methods of supervised classifications were K-Nearest Neighbor and Naïve Bayes. The steps for implementing these two methods were mostly the same with small differences in implementation as the figures (1 and 2) show.

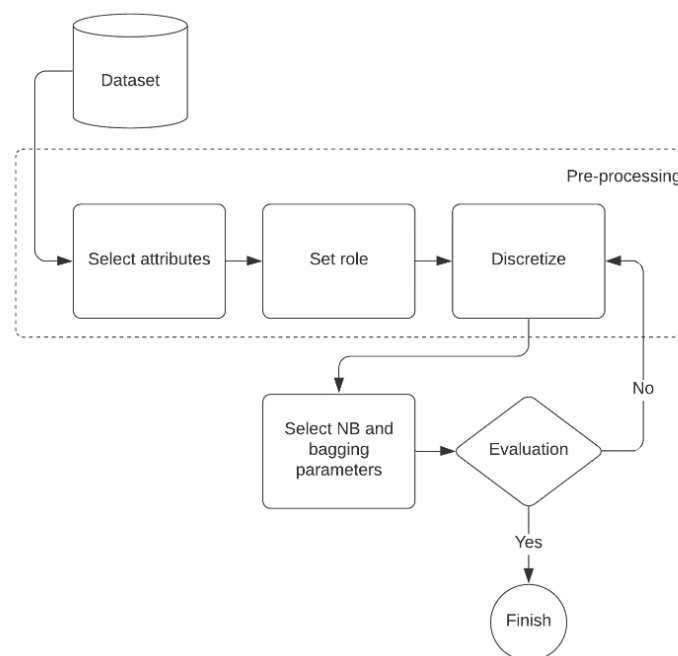


Figure 3 NB implementation process

4.2.1 Data preparation and Pre-processing

The following steps need to be performed for data preparation as it can be seen from the “pre-processing” part of figure 1 and 3.

4.2.1.1 Missing/Inconsistent values

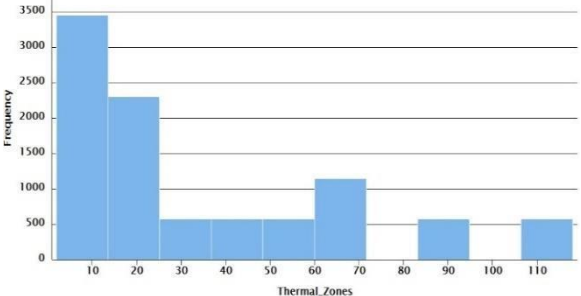
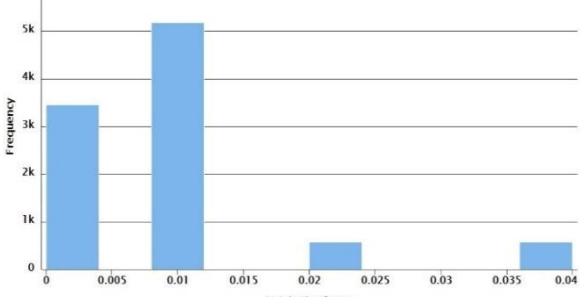
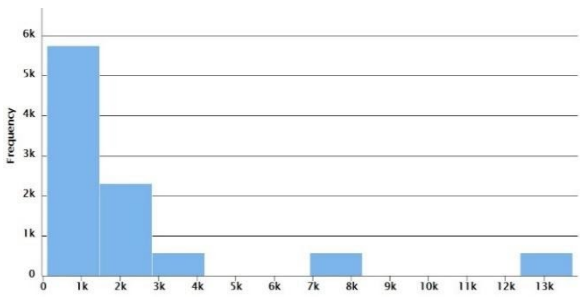
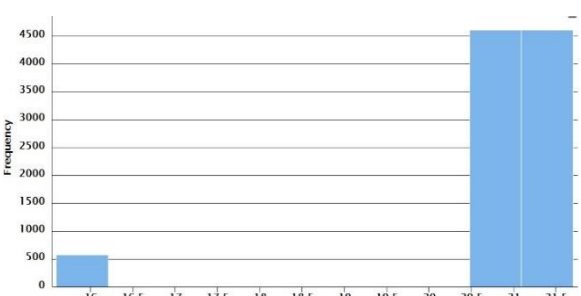
There is neither missing nor inconsistent values in the dataset.

4.2.1.2 Outliers

Based on the dataset statistics, there are some values in some attributes which were considered as outliers. The key reason for setting some parts of the values aside was the gaps between the amounts. The “Original Bar Chart” column in table 2 shows the initial spread of the values.

Table 2 Filtered Attributes

Filtered Attributes	Applied Range		Original Bar chart
	Min	Max	
Gross_floor_area_GFA_(ft^2)	2500	251k	
Roof_footprint_(ft^2)	259	78k	
Wall_Area_(ft^2)	1721	53k	

Thermal_Zones	2	72	 <table><caption>Frequency Distribution of Thermal_Zones</caption><thead><tr><th>Thermal_Zones</th><th>Frequency</th></tr></thead><tbody><tr><td>10</td><td>3400</td></tr><tr><td>20</td><td>2200</td></tr><tr><td>30</td><td>500</td></tr><tr><td>40</td><td>500</td></tr><tr><td>50</td><td>500</td></tr><tr><td>60</td><td>1100</td></tr><tr><td>70</td><td>0</td></tr><tr><td>80</td><td>0</td></tr><tr><td>90</td><td>500</td></tr><tr><td>100</td><td>0</td></tr><tr><td>110</td><td>500</td></tr></tbody></table>	Thermal_Zones	Frequency	10	3400	20	2200	30	500	40	500	50	500	60	1100	70	0	80	0	90	500	100	0	110	500						
Thermal_Zones	Frequency																																
10	3400																																
20	2200																																
30	500																																
40	500																																
50	500																																
60	1100																																
70	0																																
80	0																																
90	500																																
100	0																																
110	500																																
Height/Roof_area	0	0.013	 <table><caption>Frequency Distribution of Height/Roof_area</caption><thead><tr><th>Height/Roof_area</th><th>Frequency</th></tr></thead><tbody><tr><td>0</td><td>3400</td></tr><tr><td>0.005</td><td>0</td></tr><tr><td>0.01</td><td>5100</td></tr><tr><td>0.015</td><td>0</td></tr><tr><td>0.02</td><td>500</td></tr><tr><td>0.025</td><td>0</td></tr><tr><td>0.03</td><td>0</td></tr><tr><td>0.035</td><td>0</td></tr><tr><td>0.04</td><td>500</td></tr></tbody></table>	Height/Roof_area	Frequency	0	3400	0.005	0	0.01	5100	0.015	0	0.02	500	0.025	0	0.03	0	0.035	0	0.04	500										
Height/Roof_area	Frequency																																
0	3400																																
0.005	0																																
0.01	5100																																
0.015	0																																
0.02	500																																
0.025	0																																
0.03	0																																
0.035	0																																
0.04	500																																
Volume/Thermal_Zone	173.27	4.2k	 <table><caption>Frequency Distribution of Volume/Thermal_Zone</caption><thead><tr><th>Volume/Thermal_Zone</th><th>Frequency</th></tr></thead><tbody><tr><td>0</td><td>5600</td></tr><tr><td>1k</td><td>2200</td></tr><tr><td>2k</td><td>500</td></tr><tr><td>3k</td><td>0</td></tr><tr><td>4k</td><td>0</td></tr><tr><td>5k</td><td>0</td></tr><tr><td>6k</td><td>0</td></tr><tr><td>7k</td><td>500</td></tr><tr><td>8k</td><td>0</td></tr><tr><td>9k</td><td>0</td></tr><tr><td>10k</td><td>0</td></tr><tr><td>11k</td><td>0</td></tr><tr><td>12k</td><td>0</td></tr><tr><td>13k</td><td>500</td></tr></tbody></table>	Volume/Thermal_Zone	Frequency	0	5600	1k	2200	2k	500	3k	0	4k	0	5k	0	6k	0	7k	500	8k	0	9k	0	10k	0	11k	0	12k	0	13k	500
Volume/Thermal_Zone	Frequency																																
0	5600																																
1k	2200																																
2k	500																																
3k	0																																
4k	0																																
5k	0																																
6k	0																																
7k	500																																
8k	0																																
9k	0																																
10k	0																																
11k	0																																
12k	0																																
13k	500																																
Heating_Setpoint_Temp	20	21.7	 <table><caption>Frequency Distribution of Heating_Setpoint_Temp</caption><thead><tr><th>Heating_Setpoint_Temp</th><th>Frequency</th></tr></thead><tbody><tr><td>16</td><td>500</td></tr><tr><td>16.5</td><td>0</td></tr><tr><td>17</td><td>0</td></tr><tr><td>17.5</td><td>0</td></tr><tr><td>18</td><td>0</td></tr><tr><td>18.5</td><td>0</td></tr><tr><td>19</td><td>0</td></tr><tr><td>19.5</td><td>0</td></tr><tr><td>20</td><td>0</td></tr><tr><td>20.5</td><td>4500</td></tr><tr><td>21</td><td>4500</td></tr><tr><td>21.5</td><td>0</td></tr></tbody></table>	Heating_Setpoint_Temp	Frequency	16	500	16.5	0	17	0	17.5	0	18	0	18.5	0	19	0	19.5	0	20	0	20.5	4500	21	4500	21.5	0				
Heating_Setpoint_Temp	Frequency																																
16	500																																
16.5	0																																
17	0																																
17.5	0																																
18	0																																
18.5	0																																
19	0																																
19.5	0																																
20	0																																
20.5	4500																																
21	4500																																
21.5	0																																

4.2.1.3 Attribute Selection

Among fifty-seven attributes in the dataset, the ID related attributes, which do not affect the results, are excluded in the first place. KNN needs a binominal target attribute for the prediction, “EUI Classification (0: better than baseline; 1: worse than baseline)” and “NPW Classification (1: better than baseline, 0: worse than baseline)” were selected for voting.

In NB classification, there were two sets of selections. The first one, as table 3 shows, was for attributes with the correlations of less than 0.75. To increase the performance of the NB classification, the second set of selection obtained by using the “Optimize Selection” parameter. Table 4 shows the second set of features.

It is good to mention that by adding other features to the second selection, the performance of NB will be decreased. The reason is that the additional attributes still have some correlation, so NB will be confused, but indeed, in reality, it is hard to find attributes without any correlations, so the first set of selection is more near the reality.

4.2.1.4 Data Type Modification

KNN classification desires numerical data to process, so “Nominal to Numerical” parameter was used to change the type of non-numerical attributes to numerical by using the “dummy coding” as the coding type, which turns every different value of the non-numerical attributes to a single attribute. The next part is to normalize the values since, for measuring the nearest distances, all the values should be on the same scale. “Z-transformation” was used for the method of normalizing, which brings every value between “-1” and “+1”. Naïve Bayes does not need the data type to be modified and normalized.

4.2.1.5 Correlated Attributes

NB tends to adopt the independent features for more efficient classification, so the attributes with the correlation of 0.75 and above were removed from the attribute selection. Table 3 shows the selected attributes in this case. KNN does not need the correlated attributes to be removed.

Table 3 First set of selection for NB

Selected Attributes for NB with the Correlation of less than 0.75	
Stories	SHGC
Gross_floor_area_GFA_(ft^2)	WWR_South
Stories/Roof_area	WWR_North
Height/Roof_area	WWR_East
Stories/Wall_area	WWR_West
Height/Wall_area	R-value_Roof
Volume/Wall_area	R-value_Wall
South_Wall_%	LPD_Reduction
East_Wall_%	Main_Building_Type
Volume/Thermal_Zone	Orientation
Heating_thermostat_Sch(hr/week)	HVAC_System
U-value_(W/(m2K))	

Table 4 Second set of selection for NB (efficient attributes)

Selected Attributes for NB - Efficient Attributes	
Building_Volume_(m ³)	WWR_East
Height_(m)	WWR_South
HVAC_System	WWR_West
Wall_area/Roof_area	

4.2.1.6 Discretize

as NB works for Nominal values, discretization helps it to be more effective and have higher performance, so the “Discretize by Binning” parameter was used in the process with 55 number of bins for the values of the features which came of the optimization operator and a series of trial and error.

4.2.2 Modelling:

The modellings of the KNN and NB methods are explained as follows.

4.2.2.1 K-Nearest Neighbor

KNN is a lazy learning type of supervised classification. The first step of modelling is to train k set of examples and then compare them with an unknown example set to find the k closest training data. The distance of closeness in the n-dimensional (number of attributes) space can be measured by proximity measures as; Nominal, Numerical, Mixed measures and Bergman Divergences types through Rapidminer.

To find the best “k” in “K-NN” parameter, “Optimize” parameter was used based on two main criteria, accuracy and recall. Recall parameter shows the actual portion of correct EUI predictions out of what the model predicted as true positives, so the highest value for recall is the primary aspect of finding the best “k”.

As table 5 shows, the highest values of recall relate to k=6,8 and 10. Between these “k” values, k=6 has the highest accuracy, so it suits the model. Also, the “weighted vote” option in the “KNN” parameter takes account of the distances between data points for prediction.

After finding the best k, the next phase is to select the measure type for finding the nearest neighbours. The “Mixed Euclidean distance” was set as the measure type since there were both types of numerical and binominal for prediction.

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!

Table 5 Sorted amounts of k by Recall using Optimize operator

Iteration	k-NN.k	Recall for positive class (=0)	Accuracy
5	6	98.81	0.968171
7	8	98.81	0.963542
9	10	98.81	0.957176
1	2	98.78	0.976852
4	5	98.72	0.962963
3	4	98.67	0.975116
11	12	98.67	0.956597
8	9	98.63	0.953125
6	7	98.54	0.956597
13	14	98.47	0.950231

4.2.2.2 Naïve Bayes

Naïve Bayes is an eager learning type of supervised classification which is low-variance and high-biased. It can work very well with even small datasets. NB is a simple probabilistic classifier that generates probabilities for different classes and selects predicted values with the highest performance for the example set. As in the theorem of NB, features have no dependencies, in the selection of attributes, this point should be taken into consideration. The implementation of NB was done by using “Naïve Bayes” modelling parameter through RapidMiner.

As there were no missing records in the dataset, there was no need to select Laplace correlation, which assigns a small default probability for those ones with the zero probabilities.

4.2.3 Evaluation

4.2.3.1 K-Nearest Neighbor

For assessing the performance of the K-NN model, the “Binominal Classification Performance” operator was applied since the target attributes are binary attributes with amounts of “0” and “1”.

Accuracy, the area under the curve (AUC), precision, recall, lift, f measure, sensitivity and specificity were the criteria for this part.

As figure 4 shows, the “Cross Validation” operator with the stratified type of sampling and 16 folds was used to evaluate the statistical performance of the training set.

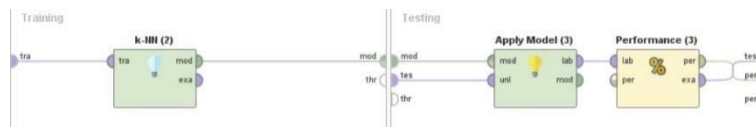


Figure 4 KNN performance evaluation in the Rapidminer

4.2.3.2 Naïve Bayes

For evaluating the performance of NB same procedure as KNN was performed.

4.3 Support Vector Machine

Another method to classify the dataset is SVM to provide a prediction for a given data based on a non-probabilistic binary linear classification. SVM is a supervised learning model that uses Java operations for improving the accuracy and pace of classification by providing different loss functions such as linear and quadratic loss functions.

The steps of implementing SVM are quite similar to data preparation of KNN method as it is shown in figure 2 which briefly mentioned in the form of CRISP-DM procedure as below;

4.3.1 Data preparation and Pre-processing

The data preparation for the SVM method is explained below.

4.3.1.1 Setting aside the outliers

same as the KNN process.

4.3.1.2 Attribute selection

the attributes selection was the same as the KNN process, and “EUI classification” and “NPW Classification” are labelled as the binominal target values.

4.3.1.3 Data modification

using “Nominal to Numerical” operator to change attribute values to numerical as dummy coding. Also, normalizing values by the “Z-transformation” technique to have a logical classification.

4.3.2 SVM Modelling

Examples in SVM are represented as points in an infinite-dimensional space (hyperplane), and by measuring the gaps between the points, different categories could be recognized. SVM can train algorithms to assign each example to one of those categories. Generally, the more distance between the test sets and training set will reduce the generalization error of classification. In order to make the separation easier, SVM maps the existing finite-dimensional space by defining kernel function $k(x, y)$ to a higher-dimensional space.

There are different types of kernel functions, such as; dot, radial, polynomial and multiquadric. As the training set in this project is normalized, the polynomial kernel was used to have more accurate results. The polynomial kernel defined as formula 1 below in the Rapidminer:

$$k(x,y)=(x*y+1)^d$$

Formula 1

“d”: kernel degree

There is another indicator defined in the Rapidminer called “Complexity constant (C)” to tolerate with misclassifications. As the “C” values increase, there will be softer boundaries, but it may cause overfitting. Different amounts for “C” were tested to avoid overfitting and over-generalization, and C=10 was selected for this purpose.

The “Convergence epsilon” and “Max iterations” remained as their default values in RM, 0.001 and 100,000, respectively.

4.3.3 Evaluation

To better distinguish the results of different methods used for classification in the same conditions, a similar procedure as the KNN method was selected for evaluating the performance of the SVM model. Therefore, the “Binominal Classification Performance” operator was applied in the “Cross-Validation” parameter, and the same criteria as KNN process were selected for measuring the outputs: accuracy, the area under the curve (AUC), precision, recall, lift, f measure, sensitivity, and specificity.

Chapter 5

5. Implementation and Results

The implementation process for the four mentioned methods in the Rapidminer, the following steps shall proceed.

5.1 Artificial Neural Network

The data is prepared according to the describes procedures, and dummy coding was used to change the format of categorical values into numerical. After data preparation, using the Cross-Validation, the ANN model was generated and evaluated. This process was repeated ten times (10-fold cross validation) in which every time a model was trained using 90% of the dataset as the training set, and 10% as the test set and the average estimation of these ten repetitions is produced as a single output.

Tables 2 and 3 contain several measures of goodness to assess the performance of the Neural Network model for separate positive classes. The capacity of the model to classify data points with zero target values is higher than the other class. However, both classes can be classified with acceptable accuracy. Figure 3 also illustrates AUC and ROC to have a better overview of the ability of the model to distinguish the target attribute classes.

Table 6 Neural Network performance results for EUI classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	96.94%	96.95%	92.39%	52.49%	119.07%	38.05%	82.92%
	0		96.94%	98.91%				

Table 7 Neural Network performance results for NPW classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	97.34%	96.62%	97.72%	97.49%	184.13%	97.00%	97.72%
	0		97.98%	97.00%				

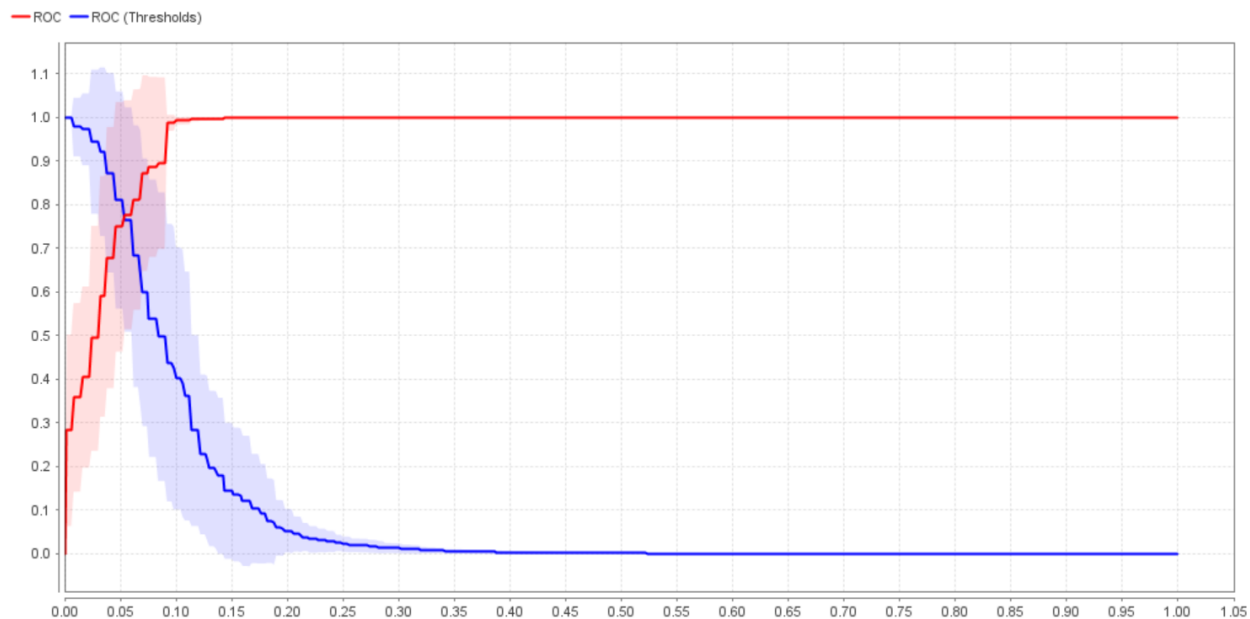


Figure 5 Neural Network area under curve and receiver operating characteristic (ROC) for EUI classification

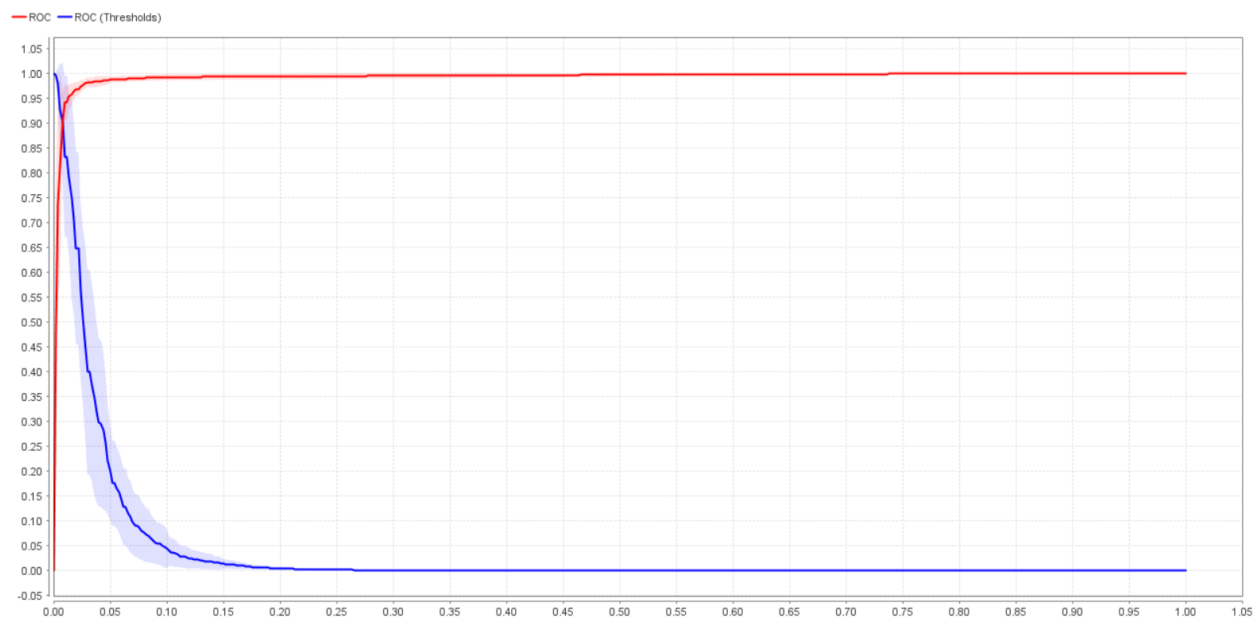


Figure 6 Neural Network area under curve and receiver operating characteristic (ROC) for EUI classification

5.2 K-Nearest Neighbor

The performance criteria of the trained KNN classification are listed below:

Table 8 KNN performance results for EUI classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	96.72%	95.04%	93.23%	97.70%	137.1%	98.16%	93.23%
	0		97.25%	98.16%				

Table 9 KNN performance results for NPW classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	96.83%	96.61%	96.61%	97.02%	182.33%	97.06%	96.76%
	0		97.02%	97.02%				



Figure 7 KNN area under curve and receiver operating characteristic (ROC) for EUI classification

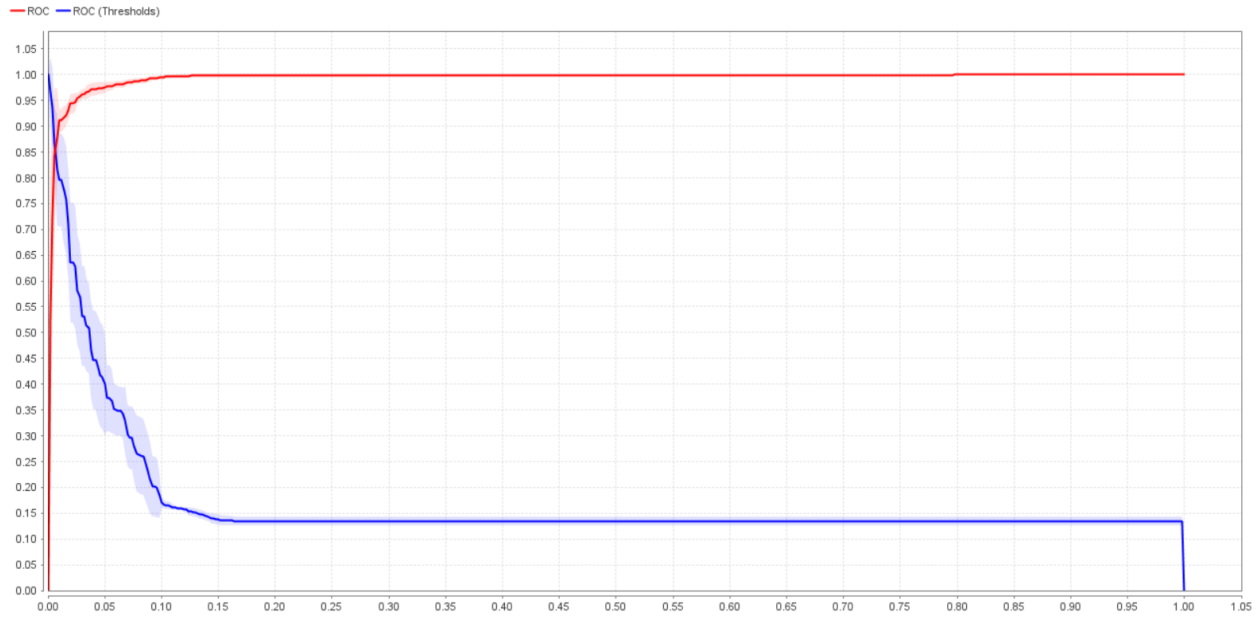


Figure 8 KNN area under curve and receiver operating characteristic (ROC) for BCR classification

5.3 Naïve Bayes

The performance criteria of the trained NB classification are listed below:

5.3.1 First Set of Selected Attributes

The results for the first set of selected attributes are shown in tables 17 and 18.

Table 10 NB performance results for EUI classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	80.06%	54.34%	86.63%	85.74%	123.77%	78.07%	86.64%
	0		95.10%	78.07%				

Table 11 NB performance results for NPW classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	75.94%	70.55%	83.37%	75.39%	155.22%	69.40%	83.37%
	0		82.60%	69.40%				

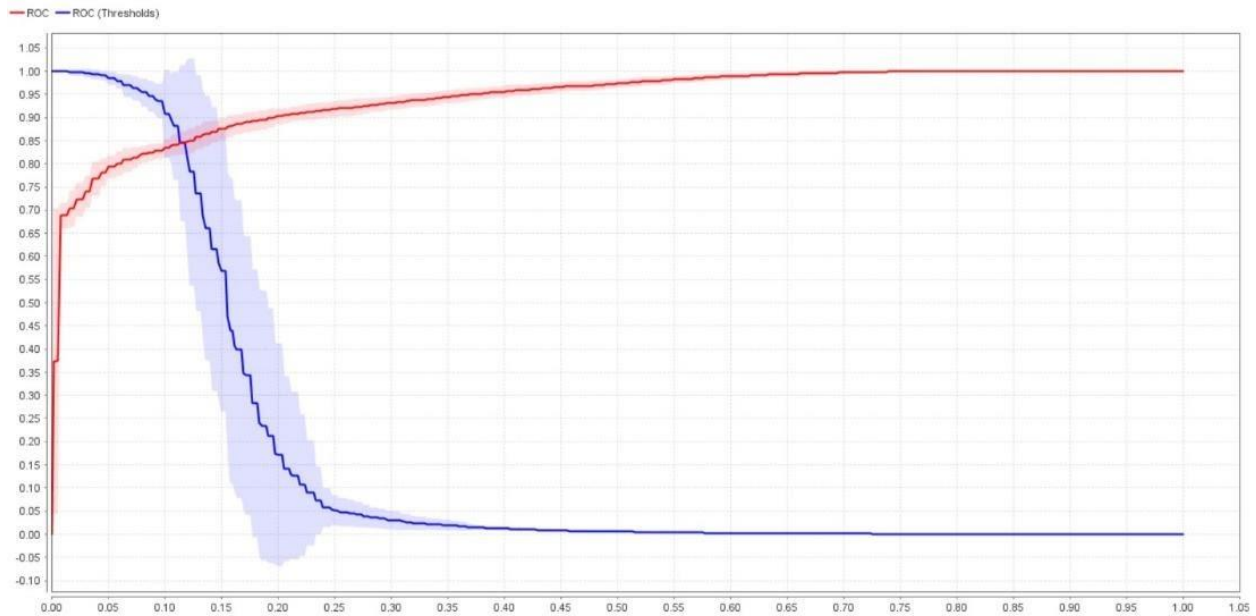


Figure 9 NB area under curve and receiver operating characteristic (ROC) for EUI classification - First set of attributes

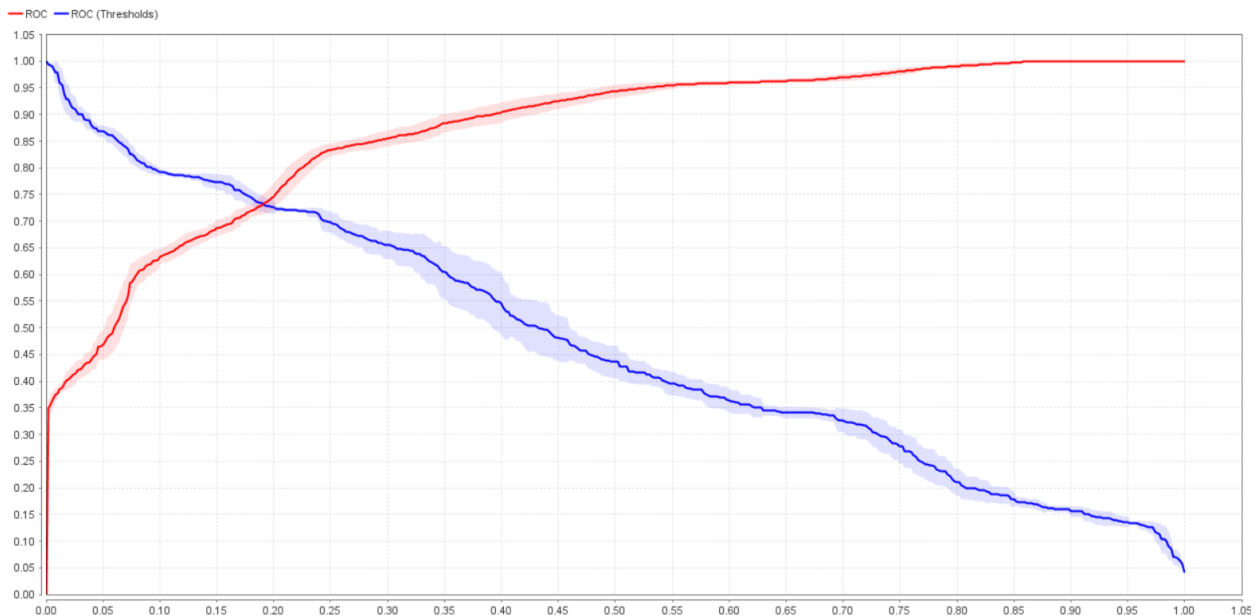


Figure 10 NB area under curve and receiver operating characteristic (ROC) for EUI classification - Second set of attributes

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!

5.3.2 Second Set of Selected Attributes

The results for the first set of selected attributes are shown in tables 17 and 18.

Table 12 NB performance results for EUI classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	86.98%	76.24%	63.56%	91.74%	116.55%	94.03%	63.57%
	0		89.55%	94.03%				

Table 13 NB performance results for NPW classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	64.47%	61.28%	57.14%	68.57%	121.41%	70.47%	57.14%
	0		66.78%	70.47%				

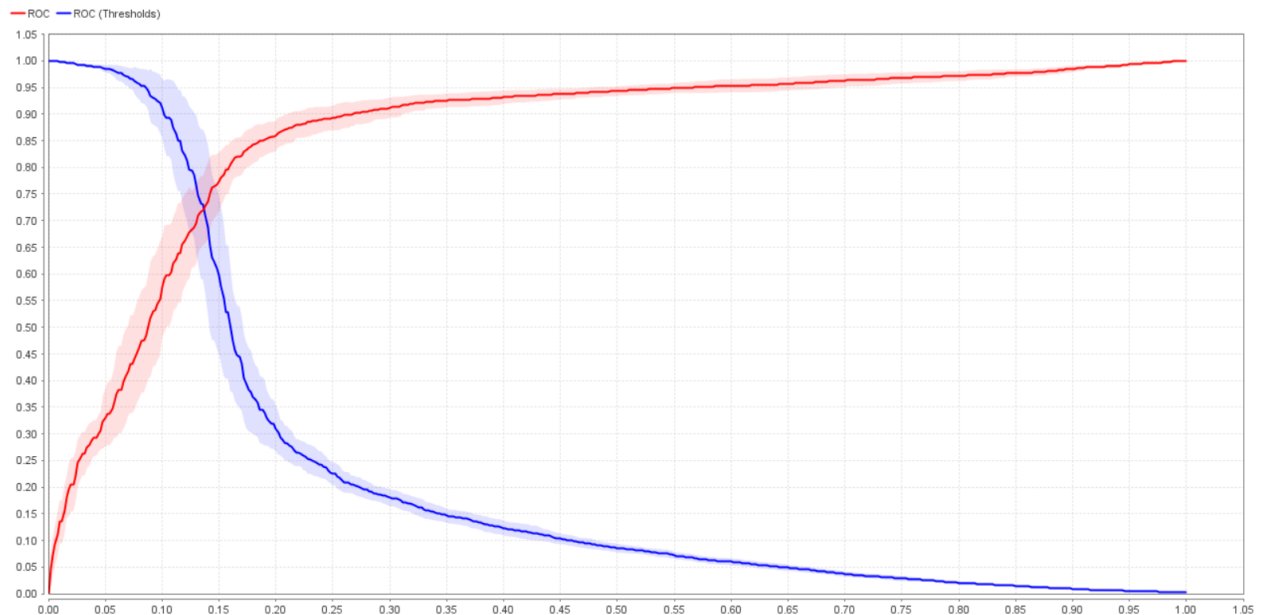


Figure 11 NB area under curve and receiver operating characteristic (ROC) for BCR classification - First set of attributes

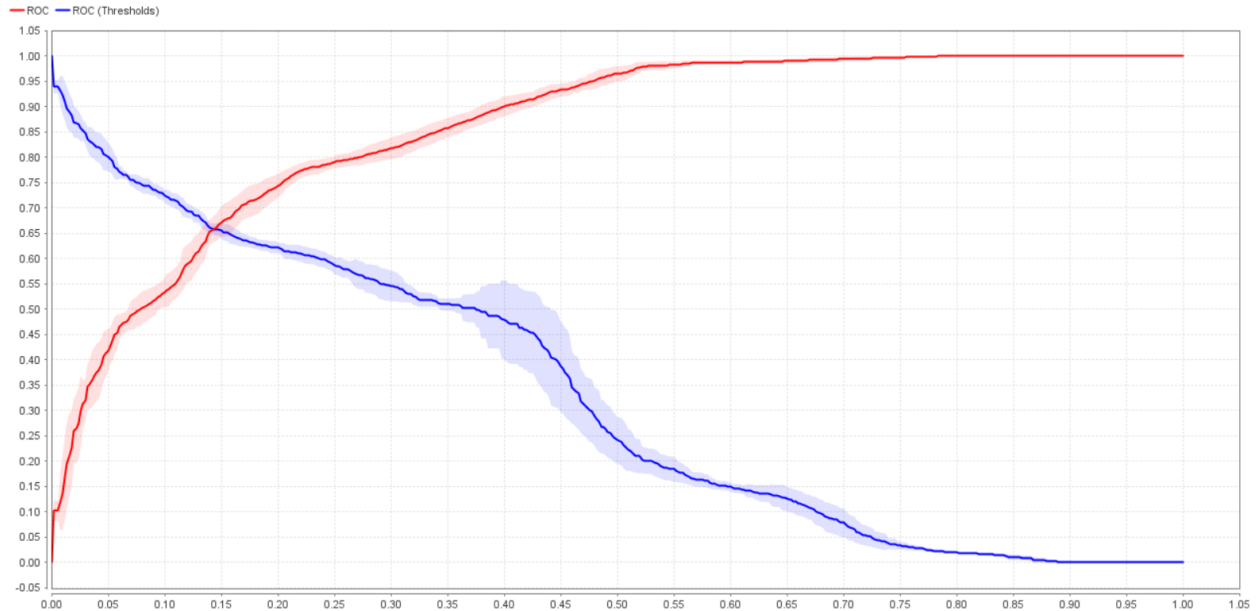


Figure 12 NB area under curve and receiver operating characteristic (ROC) for BCR classification - Second set of attributes

Even though better results were expected for efficient attributes, the accuracy of the model decreased a little. It could be found out the more attributes involved in the NB model, the higher accuracy will be captured.

5.4 Support Vector Machine

The main attributes weight table generated by the SVM parameter illustrated below table.

Table 14 SVM most effective attributes sorted weight table

Attribute	Weight
LPD_Reduction	0.533251443
R-value_Roof	0.268709853
HVAC_System = AEDG_Office_HVAC_VAV_DX_Coil	0.17120573
HVAC_System = AEDG_Office_HVAC_Fan_Coil_DOAS	0.142377764
R-value_Wall	0.104082603
HVAC_System = AEDG_Office_HVAC_ASHP_DOAS	0.095060826
Orientation = 180	0.077397373
HVAC_System = AEDG_K12_HVAC_GSHP_DOAS	0.077016546
HVAC_System = AEDG_Office_HVAC_WSHP_DOAS	0.06740905
HVAC_System = AEDG_Office_HVAC_VAV_Chilled_Water	0.05727294
Building_Model = Adjusted - ASHRAE9012013MediumOfficeDetailed	0.054564327
Lighting_Type = Troffer	0.049406383
Cooling_Setpoint_Temp	0.045038509
North_Wall_%	0.043043769
South_Wall_%	0.040750166
HVAC_System = AEDG_K12_Fan_Coil_DOAS	0.040125687

Thermal_Zones	0.035836455
Building_Model = Adjusted - ASHRAE9012013RetailStripmall	0.033678733
U-value_(W/(m2K))	0.032938795
East_Wall_area(m^2)	0.031041649
West_Wall_area(m^2)	0.031041649
Air_or_Arg_(0: air; 1: arg) = 1	0.029308871
Gross_floor_area_GFA_(ft^2)	0.022801427
Main_Building_Type = Office	0.017545411
Height_(m)	0.017233691
Roof_footprint_(ft^2)	0.017083411
Number_of_glazings(0: dbl; 1: trp) = 1	0.01577225
Building_Volume_(m^3)	0.015740145
North_Wall_area(m^2)	0.014469888
South_Wall_area(m^2)	0.013575806
Volume/Wall_area	0.012134168
Building_Model = Adjusted - ASHRAE9012013PrimarySchool	0.011943612
Main_Building_Type = School	0.011943612
Wall_Area_(ft^2)	0.010678488
Stories	0.010670538
Volume/Roof_area	0.005283025
North_Wall/Volume	0.003547619
South_Wall/Volume	0.003397609
Rise_(low_or_high) = low	0.00310507
Building_Model = Adjusted - ASHRAE9012013FullServiceRestaurant	0.002278338

The performance criteria of the trained SVM classification are listed below:

Table 15 SVM performance results for EUI classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	96.04%	92.48%	94.03%	97.20%	137.53%	96.87%	94.03%
	0		97.54%	96.87%				

Table 16 SVM performance results for NPW classification

Attribute Selection	Positive Class	Accuracy	Precision	Recall	F measure (positive class: 0)	Lift (positive class: 0)	Sensitivity (positive class: 0)	Specificity (positive class: 0)
All	1	94.25%	95.51%	92.03%	94.69%	175.35%	96.19%	92.03%
	0		93.21%	96.19%				

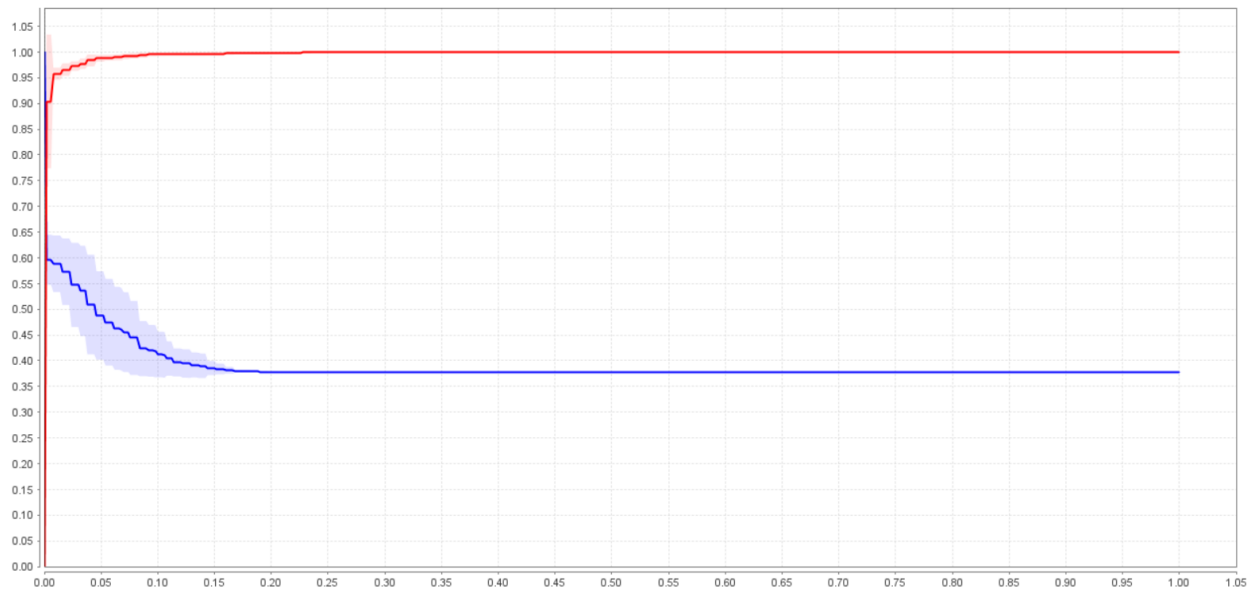


Figure 13 SVM area under curve and receiver operating characteristic (ROC) for EUI classification

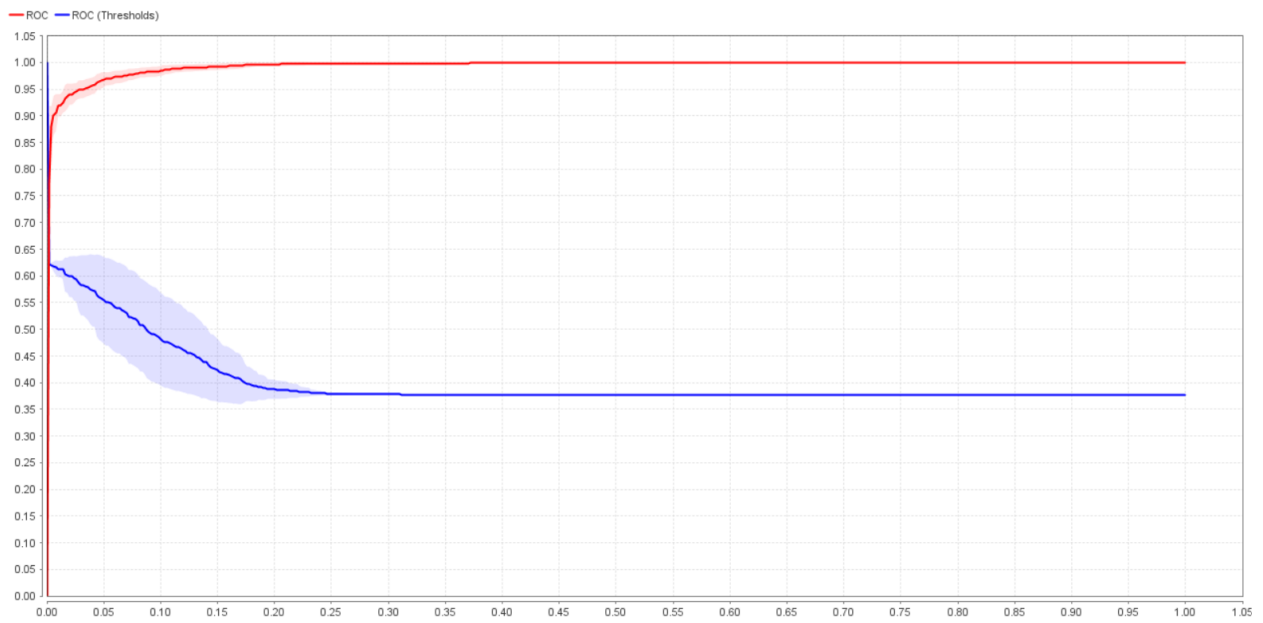


Figure 14 SVM area under curve and receiver operating characteristic (ROC) for BCR classification

Chapter 6

6. Discussion

The aim of this study is to develop a model that, based on the given attributes, can evaluate whether a building design has better or worse performance than its baseline version. Multiple different models were trained in order to find the one that performs the best based on their accuracy, precision and recall. Models looking at both EUI and BCR were separately discussed in this section.

6.1 Energy Use Intensity

Table 11 illustrates the performance results for the EUI classification.

Table 17: Modelling Performances Results for EUI Models

Method	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)
KNN	96.72%	95.40%	97.25%	93.23%	98.16%	99.37%
NB (First group attributes)	80.06%	54.34%	95.10%	86.63%	78.07%	88.09%
NB (Efficient group attributes)	86.98%	76.24%	89.55%	63.56%	94.03%	89.44%
SVM	96.04%	92.48%	97.54%	94.03%	96.87%	99.46%
ANN (Nural Network)	96.94%	96.95%	96.94%	92.39%	98.81%	97.62%

Considering accuracy as the criteria for the best model among the implemented ones, ANN has the highest figure with 96.94%. The term of accuracy shows the portion of correct classification of both classes without the capability of distinguishing 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, SVM has the highest figure with 96.04%. 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 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 as better than the baseline.

Considering recall (positive class = 0=EUI better than the baseline) as the criteria for the best model among the implemented ones, ANN has the highest figure with 98.81%.

The Area Under Curve (AUC) shows the ability of the model to distinguish the target classes. The AUCs of all models have acceptable amounts of more than 88%, and it shows all the models can differentiate between the EUI=0 or EUI=1 accurately. Figure 15 shows the performance of EUI classification based on the proximity metrics.

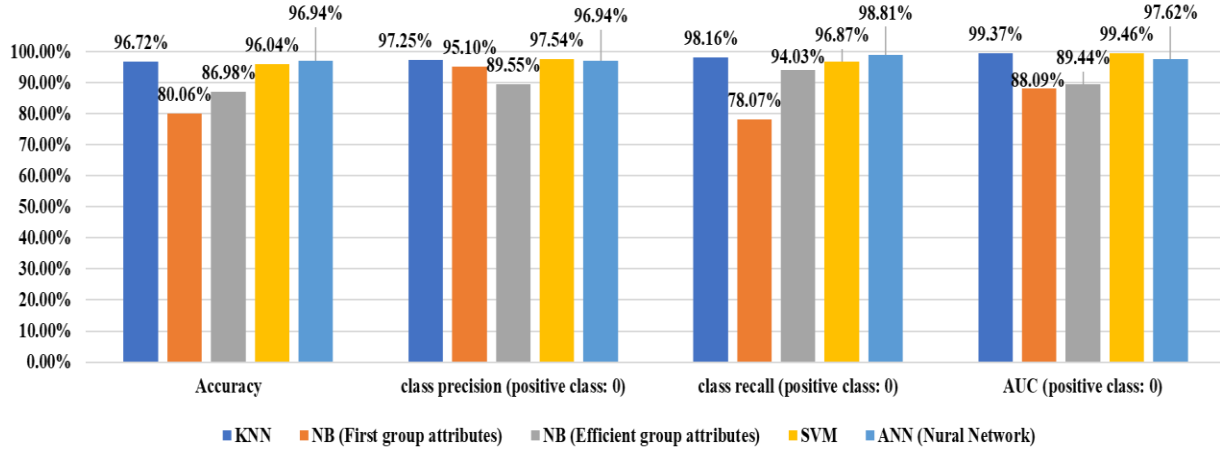


Figure 15 EUI Classification models Performance

In conclusion, in line with the aim of this study, Artificial Neural Network (ANN) provides a model with desirable performance in terms of the recall and accuracy for EUI classification between the applied classifiers.

6.2 Benefit to Cost Ratio

Similar to EUI performance results, the BCR classification is performed for the mentioned proximity measures. Table 12 illustrates the performance results for BCR classification.

Table 18 Modelling Performances Results for BCR Models

Method	Accuracy	class precision (positive class: 1)	class precision (positive class: 0)	class recall (positive class: 1)	class recall (positive class: 0)	AUC (positive class: 1)
KNN	96.83%	96.61%	97.02%	96.61%	97.02%	99.30%
NB (First group attributes)	75.94%	70.55%	82.60%	83.37%	69.40%	81.78%
NB (Efficient group attributes)	64.47%	61.28%	66.78%	57.14%	70.47%	67.79%
SVM	94.25%	95.51%	93.21%	92.03%	96.19%	99.08%
ANN (Nural Network)	97.34%	96.62%	97.98%	97.72%	97.00%	98.69%

The highest accuracy is concluded for the ANN classifier with a rate of 97.34% even though the KNN classifier presented a high accuracy in the second place. The highest precision for positive class (BCR better than the baseline) results for ANN classifier

with the precision of 96.62%. For the recall measure, which is the main criteria for deciding the best classifier in this project, ANN model result in highest accuracy of 97.72%.

The AUC for all models except the NB (efficient group attributes) is higher than 81%, which similar to EUI classification, all of the models can properly distinguish the target attributes from an acceptable perspective. Figure 16 shows the performance of BCR classification based on the proximity metrics.

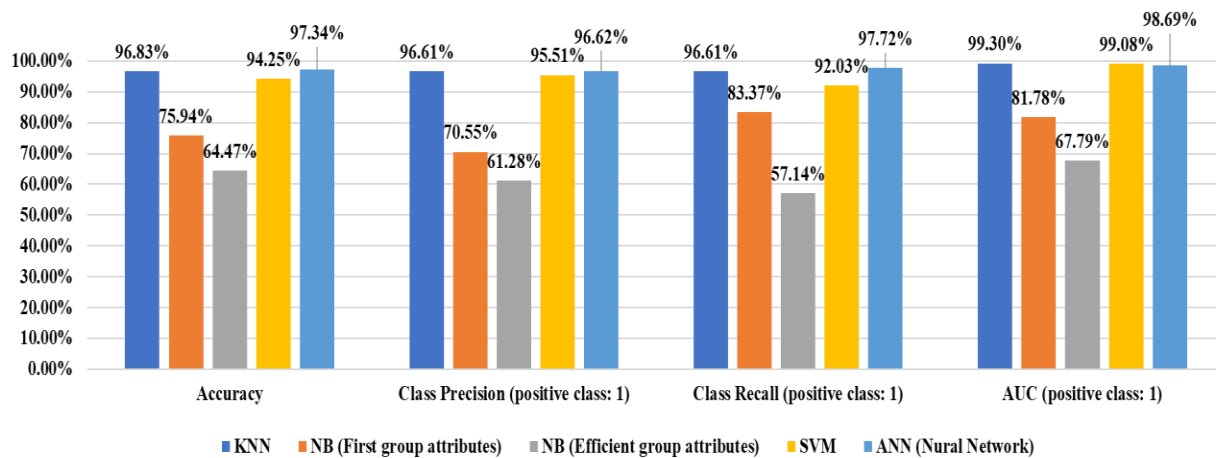


Figure 16 BCR Classification models Performance

In the conclusion of BCR classification, Artificial Neural Network (ANN) is another time selected as the best classifier between these four classification methods with the highest accuracy and recall values and the other performance metrics.

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!

Chapter 7

7. Conclusion

This study represents the classifications of different energy-related parameters of buildings in terms of EUI and BCR by using data mining approaches. The dataset is in continuation of “*Economy-energy trade-off automation – A decision support system for building design development*”[1] article, which focuses on the integration of the energy simulation and cost analysis at the early schematic stage of the design.

Four primary classification methods of KNN, NB, SVM and ANN are used for classifying the binary target attributes of EUI and BCR for different designs of buildings regarding the specified baselines. The models distinguish the designs with EUI and BCR better than the baselines and classify them accordingly. Therefore, the models can predict and evaluate whether an unknown building design is acceptable in terms of EUI and BCR or not.

In the end, the performances of the mentioned methods are compared regarding proximity measures of; accuracy, precision, recall and AUC. As the study is searching for the actual correct predicted EUI (or BCR) cases, the recall proximity measure has higher importance for comparison than the precision. Thus, the ANN model is selected as the best model with the highest recall of 98.81 and 97.98 for EUI and BCR, and the highest accuracy of 96.94 and 97.34, respectively.

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!

8. References

- [1] M. Nik-Bakht, R. O. Panizza, P. Hudon, P.-Y. Chassain, and M. Bashari, “Economy-energy trade off automation – A decision support system for building design development,” *Journal of Building Engineering*, vol. 30, p. 101222, Jul. 2020, doi: 10.1016/j.jobbe.2020.101222.
- [2] B. Yildiz, J. I. Bilbao, and A. B. 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, Jun. 2017, doi: 10.1016/j.rser.2017.02.023.
- [3] *Editorial Preface*. .
- [4] 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.
- [5] A. Saxena and A. Saad, “Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems,” *Applied Soft Computing*, vol. 7, no. 1, pp. 441–454, Jan. 2007, doi: 10.1016/j.asoc.2005.10.001.
- [6] “Operator Manual - RapidMiner Documentation.”
<https://docs.rapidminer.com/latest/studio/operators/> (accessed Aug. 11, 2020).