Energy Efficiency

MACHINE LEARNING (EE 257) PROJECT REPORT

KASHYAP MEHTA

ID: 014489891

Project Partner: Sarvil Modi

Table of Contents

1. Data Set Description

The heating load (HL) and the cooling load (CL) is required for speciation of the cooling and heating equipment when it comes to design efficient building, to maintain comfortable indoor air conditions based on user requirements. To evaluate the required heating and cooling capacities in the building, engineers and building designers need information about the characteristics of the buildings and of the inured space (for instance orientation, occupancies and dimensions). For this reason, the dataset consists of 8 input variables namely surface area, relative compactness, glazing area distribution, overall height, orientation, wall area, glazing area and, roof area to determine the output variables Heating Load and Cooling Load of residential buildings.

The dataset consists of 768 data points and 8 input features, aiming to classify two real valued responses [1]. The labels are transformed into classes by approximating each class based on range of data defined. The project looked into evaluating the cooling load and heating load requirements of building as a function of building parameters to improve the efficiency of the building and Energy analyzes using 12 different building shapes simulated in software Ecotect. The buildings differ with respect to the orientation, the glazing area distribution, and glazing area, amongst other parameters. The project converts to multi-class classification problem as it converts the responses to rounded to the nearest integer and defining classes based on the range defined.

The dataset contains eight input features and two labels. The aim is to use the eight features to predict each of the one response (Heating Load), Cooling Load is not considered as label in this project.

Specifically:

- Relative Compactness
- Roof Area
- Surface Area
- Wall Area
- Glazing Area Distribution
- Orientation
- Glazing Area
- Overall Height
- Cooling Load
- Heating Load

In this project the aim is for classification, hence heating load is only considered for classification by defining each class.

2. Data Set Visualization

Dataset is provided in csv file which is feed into the database using pandas library. The dataset consists of 8 input features with column name from X1 to X8 while the dataset also consists of two label parameters naming Y1 and Y2 as represented in Fig. 1. The dataset consists of 768 data points for training, validation and testing of the model.

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
0	0.98	514.5	294.0	110.25	7.0	2	0.0	0	15.55	21.33
1	0.98	514.5	294.0	110.25	7.0	3	0.0	0	15.55	21.33
2	0.98	514.5	294.0	110.25	7.0	4	0.0	0	15.55	21.33
3	0.98	514.5	294.0	110.25	7.0	5	0.0	0	15.55	21.33
4	0.90	563.5	318.5	122.50	7.0	2	0.0	0	20.84	28.28
...	...	...	...	...	...	...	...	...	...	...
763	0.64	784.0	343.0	220.50	3.5	5	0.4	5	17.88	21.40
764	0.62	808.5	367.5	220.50	3.5	2	0.4	5	16.54	16.88
765	0.62	808.5	367.5	220.50	3.5	3	0.4	5	16.44	17.11
766	0.62	808.5	367.5	220.50	3.5	4	0.4	5	16.48	16.61
767	0.62	808.5	367.5	220.50	3.5	5	0.4	5	16.64	16.03

Figure 1: Raw Dataset

To understand each feature input the column names have been renamed using the pandas library as represented in Fig. 2. The input features from X1 to X2 are replaced with Wall Area, Relative Compactness, Roof Area, Orientation, Surface Area, Overall Height, Glazing Area Distribution and Glazing Area respectively. While the labels Y1 and Y2 are replaced with Heating Load and Cooling Load.

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heating Load	Cooling Load
0	0.98	514.5	294.0	110.25	7.0	2	0.0	0	15.55	21.33
1	0.98	514.5	294.0	110.25	7.0	3	0.0	0	15.55	21.33
2	0.98	514.5	294.0	110.25	7.0	4	0.0	0	15.55	21.33
3	0.98	514.5	294.0	110.25	7.0	5	0.0	0	15.55	21.33
4	0.90	563.5	318.5	122.50	7.0	2	0.0	0	20.84	28.28
...	...	...	...	...	...	...	...	...	...	...
763	0.64	784.0	343.0	220.50	3.5	5	0.4	5	17.88	21.40
764	0.62	808.5	367.5	220.50	3.5	2	0.4	5	16.54	16.88
765	0.62	808.5	367.5	220.50	3.5	3	0.4	5	16.44	17.11
766	0.62	808.5	367.5	220.50	3.5	4	0.4	5	16.48	16.61
767	0.62	808.5	367.5	220.50	3.5	5	0.4	5	16.64	16.03

768 rows × 10 columns

Figure 2: Dataset with renamed Column

Heating Load	Cooling Load
16	21
16	21
16	21
16	21
21	28
...	...
18	21
17	17
16	17
16	17
17	16

Figure 3: Labels from float to integer

The aim of the project is to use classification model but the provided dataset consists of labels which are in float values, hence the labels are transformed from float to integer with nearest element rounded off. The Fig. 3 represents the label parameters after converting floating values to integer with rounded off using pandas library.

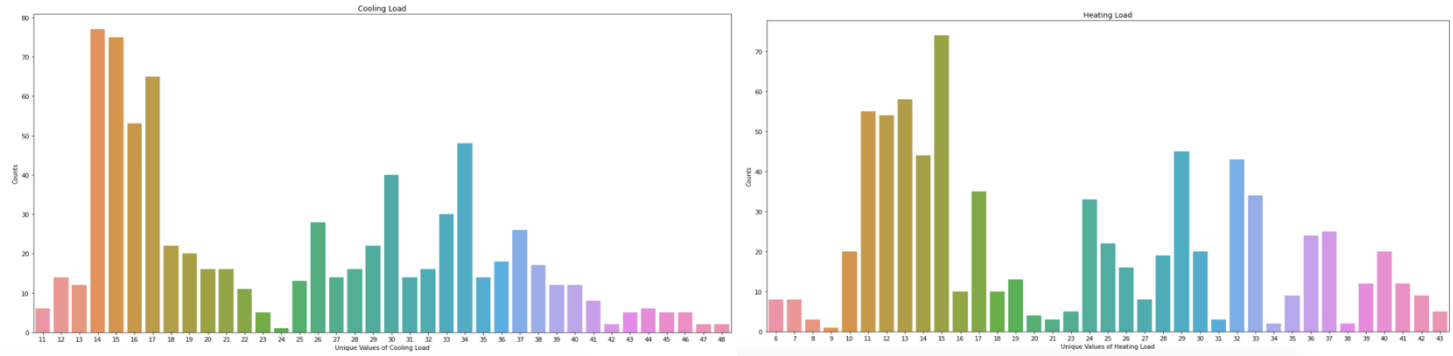

Figure 4: Labels Histogram

To find the unique quantities exists in the labels after rounding off to nearest integer, the histogram is plotting as represented in Fig. 4 using seaborn library. As depicted from the histogram of Heating Load and Cooling Load that the distribution of data is wide ranging from 6 to 43 for Heating Load while for Cooling Load it ranges from 11 to 48. As well as the maximum number of unique counts obtained is 71 for 15 value of Heating Load and 77 for 15 value of Cooling and the minim count obtained is 2 for 9 value of Heating Load and 2 for 24 value of Cooling Load. From the histogram it could be established that the number of unique integer values is very less if we use classification technique directly, hence the integers are further classified with particular ranges defines the particular class.

The table below represents the class defined for each integer value of labels:

Integer Range	Heating Load	Cooling Load
Less than 10	Class 0	Class 0
10 – 15	Class 1	Class 1
15 – 20	Class 2	Class 2
20 – 25	Class 3	Class 3
25 – 30	Class 4	Class 4
30 – 35	Class 5	Class 5
35 – 40	Class 6	Class 6
40 – 45	Class 7	Class 7
Greater than 45	Class 8	Class 8

Table 6: Classification of Labels

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000

Figure 7: Statistics of Input Features

Heating Load	Cooling Load
Class 2	Class 3
Class 2	Class 3
Class 2	Class 3
Class 2	Class 3
Class 3	Class 4
...	...
Class 2	Class 3
Class 2	Class 2
Class 2	Class 2
Class 2	Class 2
Class 2	Class 2

Figure 5: Labels in different class

The statistics of the input features are obtained using pandas library as represented in Fig. 7. It could be observed that the dataset does not consists of any vast variations and hence it does not require to remove any outliers.

Fig. 8 shows, histogram of the input features which represents that the Orientation column has only four unique values with equal distribution. The figure also helps to find the out that the distribution of data is equal for Overall Height, Surface Area and Relative Compactness. While distribution Glazing Area and Glazing Area Distribution is similar and Roof Area is having unequal distribution of data.

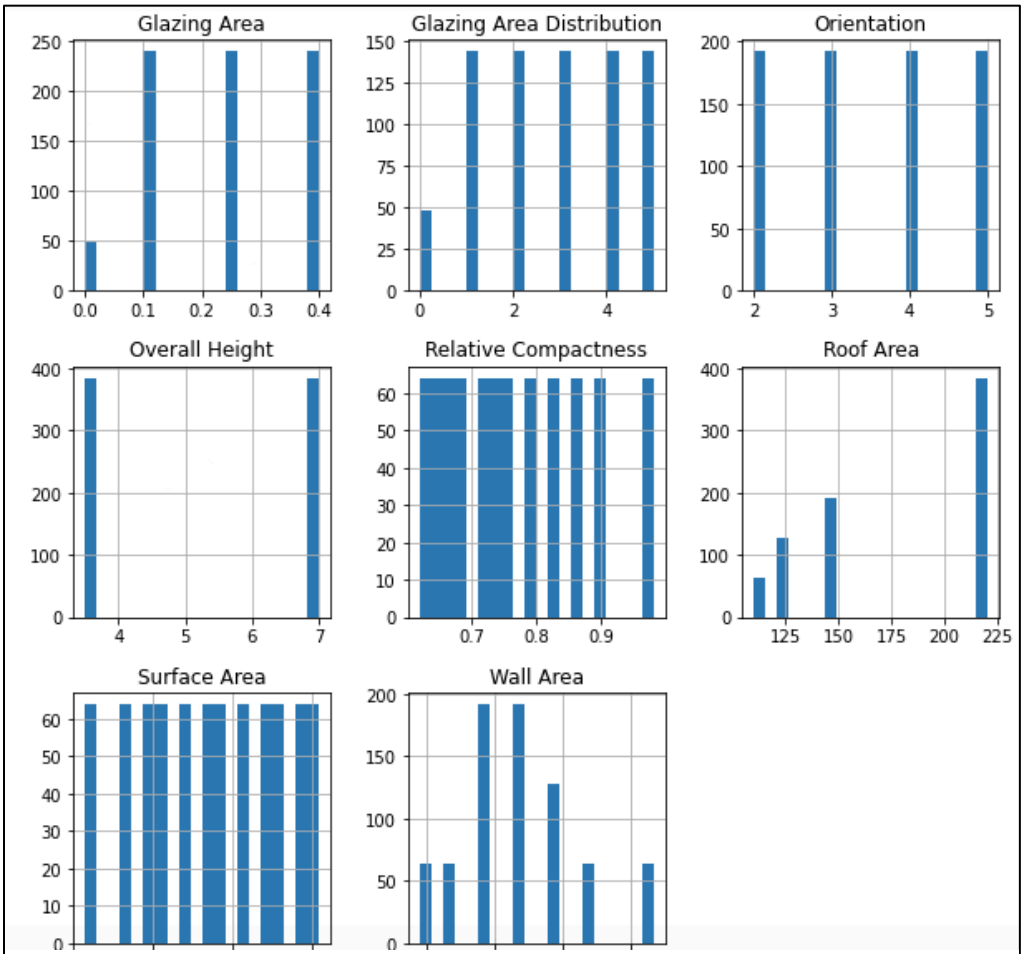

Figure 8: Histogram of Input Features

From the Fig. 9, we can see some information about correlations between all variables. For example, the Overall height (an input) has a strong correlation (0.90) with the output – Cooling Load. Besides, the pair plot depicts there is some relationship between them. For the Overall height and Cooling Load plot, there is only 2 values of overall height due to the distribution and which makes difficult to see the linear correlations of those variables. We will use preprocessing method to refine the distributions.

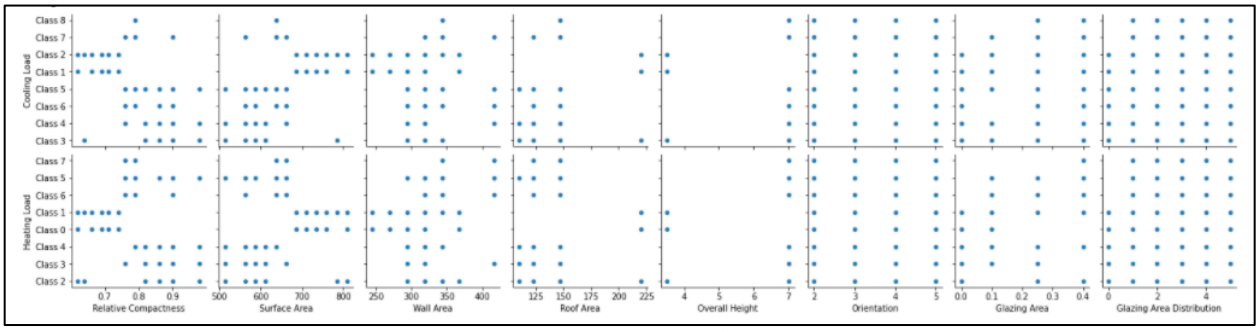

Figure 9: Dataset Correlation

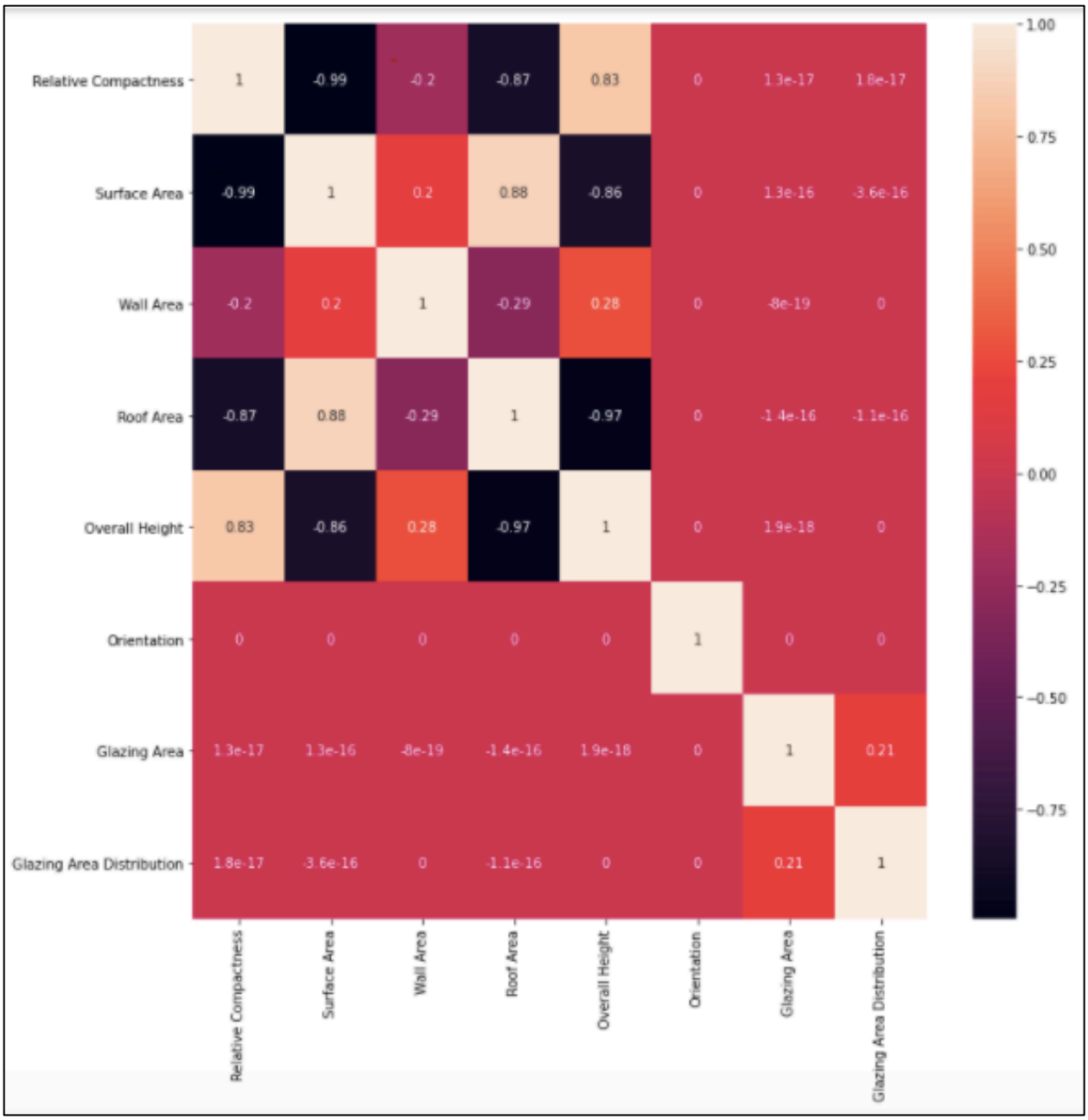

Figure 10: Correlational Matrix

To correlation of input features on the labels, the correlation matrix is plotted as shown in Fig. 10. The matrix represents that Orientation has zero correlation for all input features and hence it represents that with changing Orientation there is no effect on other input features. While the Glazing area and Glazing area distribution have some correlation among themselves. For Relative Compactness and Surface Area, there is very high amount of correlation which is near to one, that states with increase in Surface Area the Relative Compactness will increase and vice versa.

The distribution of each classes is visualized in Fig. 11, which represents that Class 1 is having the highest amount of data points in the dataset for Heating Load and Cooling Load, while Class 0 does not exist for Cooling Load. Class 8 is having least number of data points for Heating Load and Cooling Load. Due to the uneven dataset for classification, the model training would be difficult as the number of training data points are very less to train the model parameters and test the model.

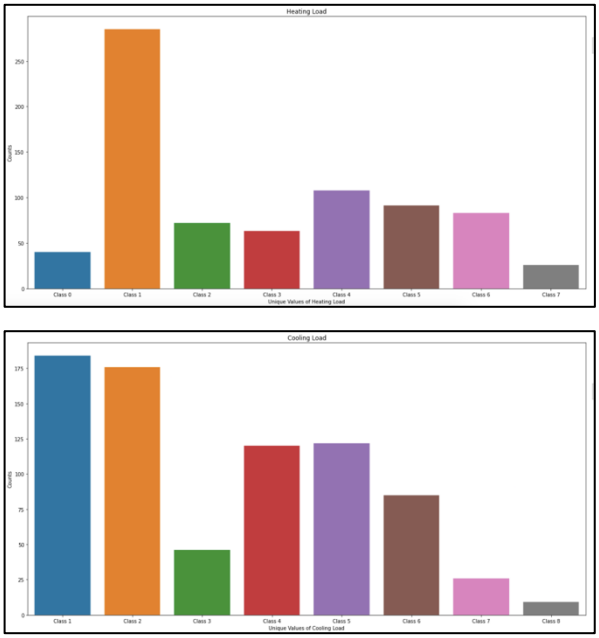

Figure 11: Labels Distribution

3. Data Set Cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Relative Compactness       768 non-null    float64
 1   Surface Area               768 non-null    float64
 2   Wall Area                  768 non-null    float64
 3   Roof Area                  768 non-null    float64
 4   Overall Height             768 non-null    float64
 5   Orientation                768 non-null    int64
 6   Glazing Area               768 non-null    float64
 7   Glazing Area Distribution  768 non-null    int64
 8   Heating Load               768 non-null    object
 9   Cooling Load               768 non-null    object
dtypes: float64(6), int64(2), object(2)
memory usage: 60.1+ KB
```

Figure 12: Dataset Information

The information of the dataset is represented in Fig. 12, all the input features consists of same number of rows (768 data points) without any null value. Hence, the dataset does not require missing value approximation or dropping the feature input points.

To verify that the dataset does not consists of any outlier box plot is used to plot for all input features as shown in Fig. 13. As represented in box plot the variance of each input features does not vary too high and hence there are not outliers in the dataset. The box plot represents the 1st Quartile and 3rd Quartile in the box while the line in the box represents the median and the end points represents the minimum and maximum values of the dataset.

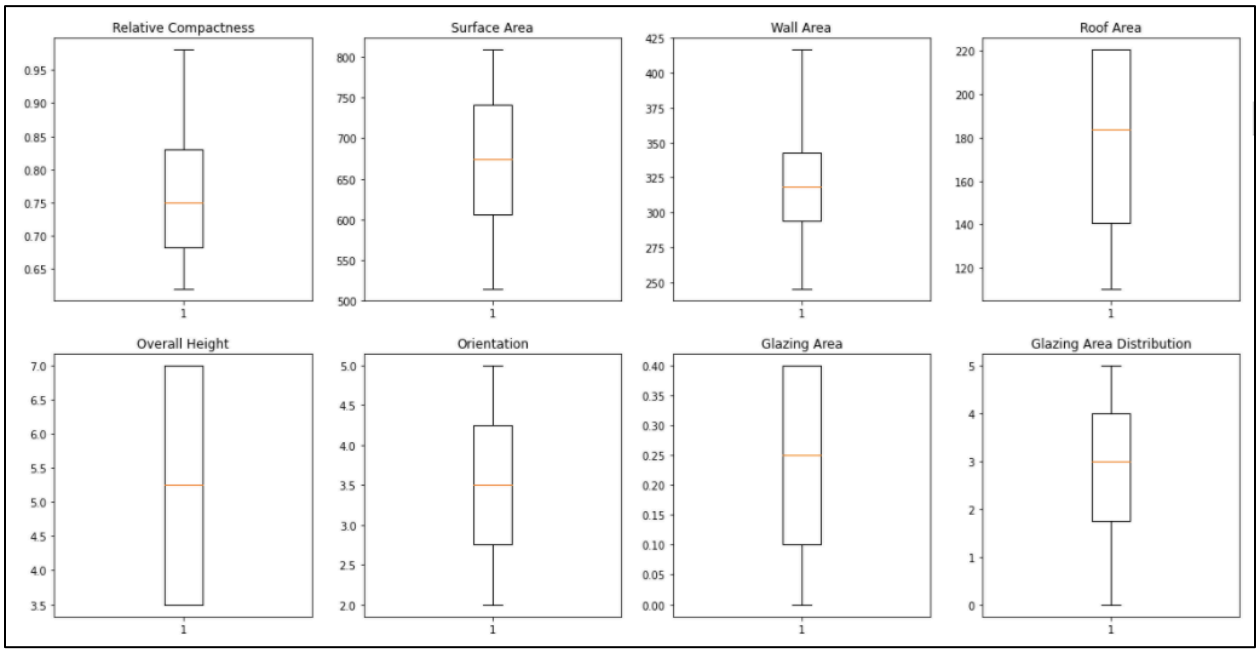

Figure 13: Box plot of Input Features

4. Related Work

Research Paper Title: "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools"

The research paper uses same dataset with eight input variables namely height, relative compactness, orientation, wall area, glazing area, surface area, roof area and glazing area distribution to predict the heating load and cooling load of the building. The research paper uses statistical and classical approach to understand the importance of each input feature over output. Research paper made comparison of classical linear regression model with the non-linear Random Forest model to predict the Heating Load and Cooling Load requirements.

The data visualization includes plotting of histogram to understand the distribution of data which helps to understand the dataset is Gaussian (Normal) Distribution or not. Normalized scatter plots for two input variables against the two output variables are also used to understand span of different ranges of values. The model uses Spearman rank correlation matrix to understand correlation between each feature with the labels as the dataset is non-Gaussian.

The research paper uses Iteratively Reweighted Least Square (IRLS) method for classical linear regression model as the baseline model. To improve the performance of the model, the research paper uses non-linear Random Forest model. To generalize the model performance, the research paper uses k-fold cross validation technique with 10 folds. To improve the statistical confidence the model is re-trained 100 times to validate the accuracy as the training and test data are spited randomly. Errors of 100 cross-validation of Random Forest model is averaged to obtain mean relative error (MRE). The result shows that the predicted output is within 2 percent of the actual

output with using random forest cross validation as model while the performance of the linear regression model was 10% variation of the predicted output as compared to the actual output.

5. Feature Extraction

To understand the importance of each feature as compared to the output feature extraction is used. One of the most frequently used models for feature extraction or feature reduction is Principle Component Analysis (PCA). In this report, PCA model is implemented and to predict the performance of the model listed models are used for classification. As represented in Fig. 10, the correlation Orientation is zero among other features which states that by increasing or decreasing the Orientation there is no effect on the output. Hence the orientation feature is dropped. By dropping Orientation input feature, the model performance has increased which proves that uncorrelated data is of no importance to the machine learning model and hence can be dropped.

To understand each feature input and to uncorrelated the correlated data, PCA is applied for 7 input features and the variance ration is represented in table below:

Feature	Variance Ratio
X1	0.87
X2	0.19
X3	2.1e-04
X4	1.4e-05
X5	1.4e-06
X6	9.1e-9
X7	7.2e-32

As represented in the table, the expected variance ratio of the all input features except Orientation which states that X1 feature is having very high importance with 0.87 variance ratio while from X3 input feature the expected variance ratio is very less and hence the importance of those features are very less. PCA uses eigenvalues and eigenvectors to transform correlated data to uncorrelated.

6. Model Development

6.1. Logistic Regression

Using Logistic Regression as the baseline model, taking 8 input features and predicting Heating Load, the performance of the model is shown in Fig. 14. The model performance is worst as if we flip a coin than also the model will able to predict 50% of the data correct. As the dataset consists of 8 classification from class 0 to class 7, the naïve algorithm will provide the probability of getting a correct result as $\frac{100}{8} = 12.5\%$. The logistic regression sill able to predict 57% data correct.

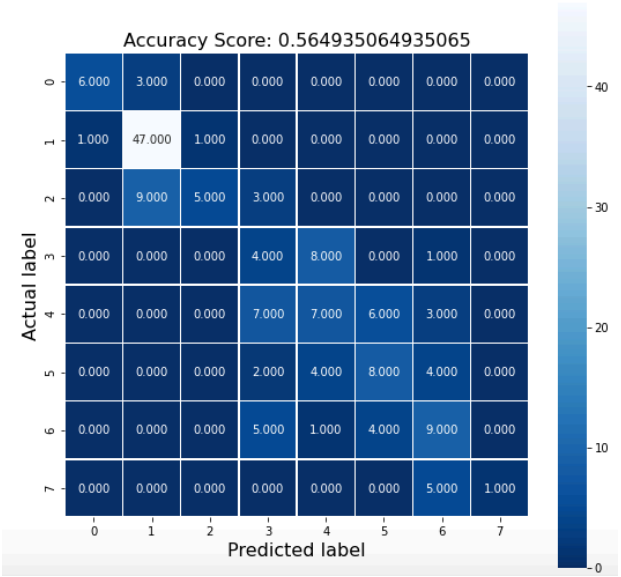

Figure 14: Confusion Matrix Logistic Regression

6.2. Quadratic Discriminant Analysis (QDA)

As the logistic regression model is not providing a good result, non-linear model is used for classification which is Quadratic Discriminant Analysis (QDA). But as shown in Fig. 15 the performance of the model obtained is worse than logistic regression which is only 31%. One of the reasons of the less performance of the model is the dataset are non-linear in nature but may not be quadratic

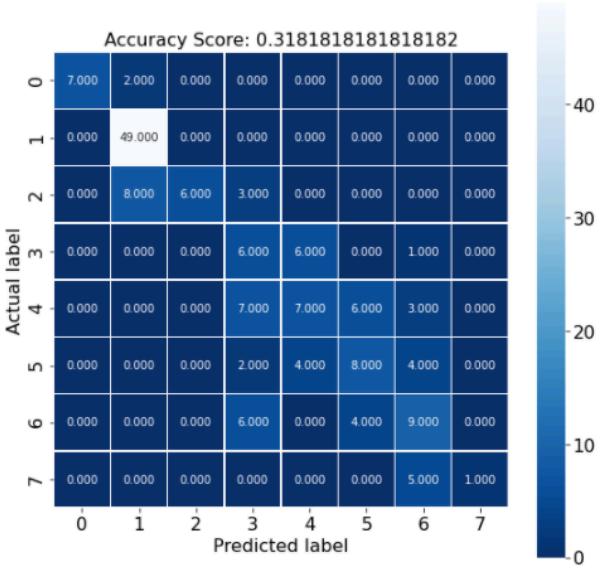

Figure 15: Confusion Matrix QDA

6.3. K – Nearest Neighbors (KNN)

To improve the performance further more non-linear model used namely k-nearest neighbors (KNN). As shown in Fig. 16, the accuracy of the model improved to 90% of the test data correctly predicted. The model designed with hyperparameter n chosen as 1 which state that the model will find the first nearest neighbor for particular data point. The mis-classification is due to the fact that the labels were approximated to nearest value and classification is defined hence the mis-classification is understood.

6.4. Support Vector Machine (SVM)

Support Vector machine is a one of the non-linear models for classification which defines decision region based on support vectors. The performance obtained using Support Vector Machine (SVM) is not that good as compared to KNN classifier as shown in Fig 17.

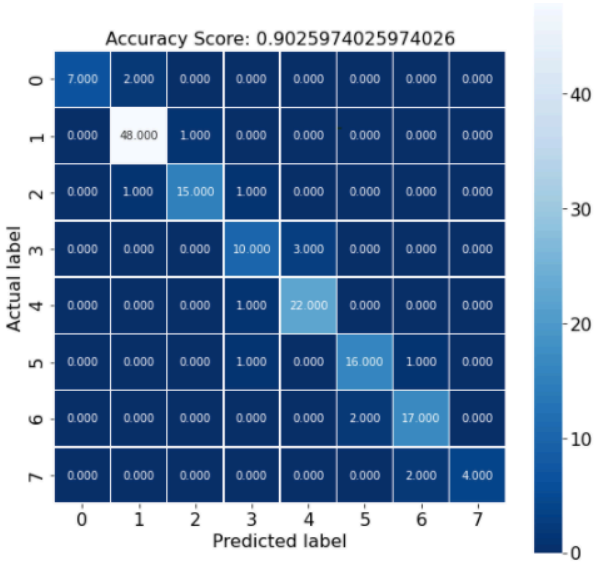

Figure 17: Confusion Matrix SVM

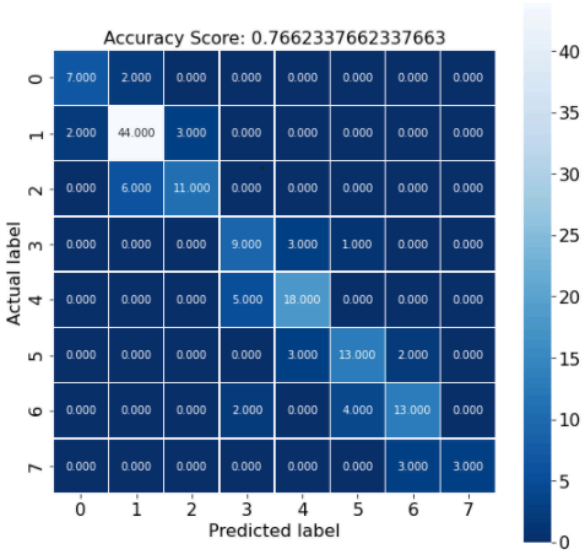

Figure 16: Confusion Matrix KNN classifier

6.5. Decision Tree

Using Decision Tree to classify heating load classes from 0 to 7 helps to predict very good result with very less data available to train and the classes defined are rounded off to the nearest hence there is high possibility of getting worst result. But with using decision tree model is able to provide 90% accuracy as represented in confusion matrix in Fig. 18.

6.6. Random Forest

Random Forest is advance version of decision tree as a greater number of trees form forest and random forest is more complex model. The performance obtained using Random Forest algorithm is 90% which is almost similar to the performance obtained for Decision Tree algorithm. The Fig 19. represents the confusion matrix of the random forest algorithm, which states that the mis-classification points are near to the defined or targeted class. For instance, for actual label 6, the predicted label obtained as 5 there are 3 data points for misclassification, as that is as expected because the heating load were available in float in the dataset while they are round to nearest integer and defined by classes with particular range. If the data point is on 15.6 it will be defined as class 2 rather which very close to class 1. So, keeping some generalization in mind, the random forest, decision tree, KNN classifier provides best results.

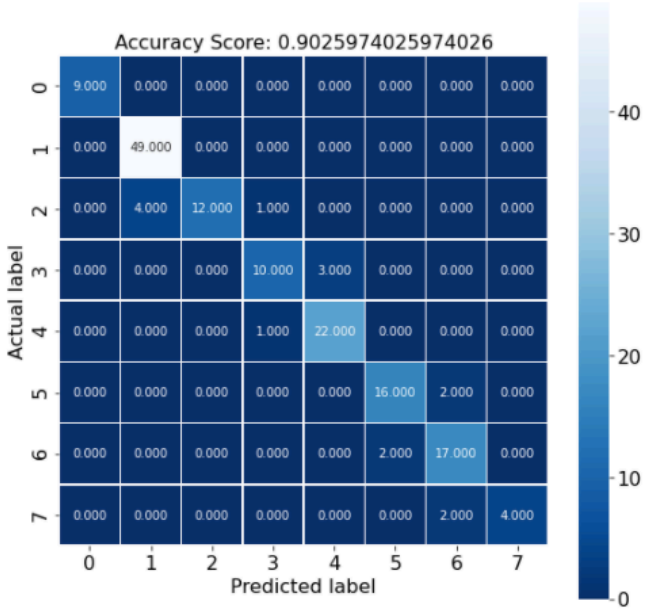

Figure 18: Confusion Matrix Decision Tree

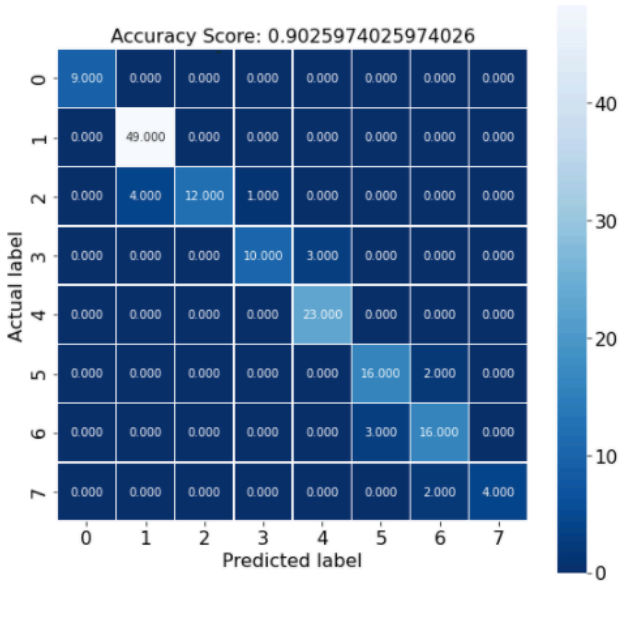

Figure 19: Confusion Matrix Random Forest

7. Model Fine tuning & Feature Set

Based on the performance of the accuracy from the confusion matrix, the random forest, decision tree and KNN classifier models provide the best result. As the models are non-linear in nature and those models provide very good response in prediction, hence it could be concluded that the dataset is non-linear and hence the non-linear models are required for prediction. QDA is one of the non-linear models but the model has quadratic nature but the dataset is not quadratic hence QDA model fails to predict and provide worst results.

7.1. Least One Out Cross Validation (LOOCV)

To increase the performance of the model the cross-validation techniques are used which will improve the training accuracy of the model as well as increase the generalization of the model. Previously, validation set approach is used which divides the dataset into training and testing by randomly splitting the dataset. But as the dataset is divided hence the training data is reduced which will reduce the training accuracy of the model. As well as the test data are randomly selected from the dataset which will reduce the generalization of model and the model will not be able to test on all dataset. Hence to overcome those issues and increase the model training performance and increase the generalization of the model Least One Out Cross Validation (LOOCV) technique is used. In LOOCV, the model is trained for n number of times where n is the total number of data points, for this report the n is 768. For each training of model one data point is assumed as test data and rest are used as training data. Taking the average of all the accuracy obtain provides the final accuracy of the model. LOOCV helps to generalize the model and increase the training accuracy. As the data points are very less hence LOOCV is used instead of k-fold cross validation.

As represented in Fig. 20, LOOCV is applied for 800 trials and the performance are obtained for all the models. As represented, the performance of the model has enhanced a lot by using LOOCV as the model is generalized with the test data and more data are used to train the model. As represented the logistic regression model previously was able to provide 56% while now the model is able to provide near to 70% accuracy. For QDA model, the performance is not increased in logarithmic as the model is quadratic in nature and the dataset is highly non-linear. Using LOOCV for Random Forest provides excellent result which is shown in Fig. 20, 95% the training accuracy is achieved. Another thing to note here is that the non-linear model like KNN, SVM, Decision tree and Random Forest achieved performance is more than 80% as the dataset is non-linear and LOOCV helps to improve the model performance and to generalize the model with the test data.

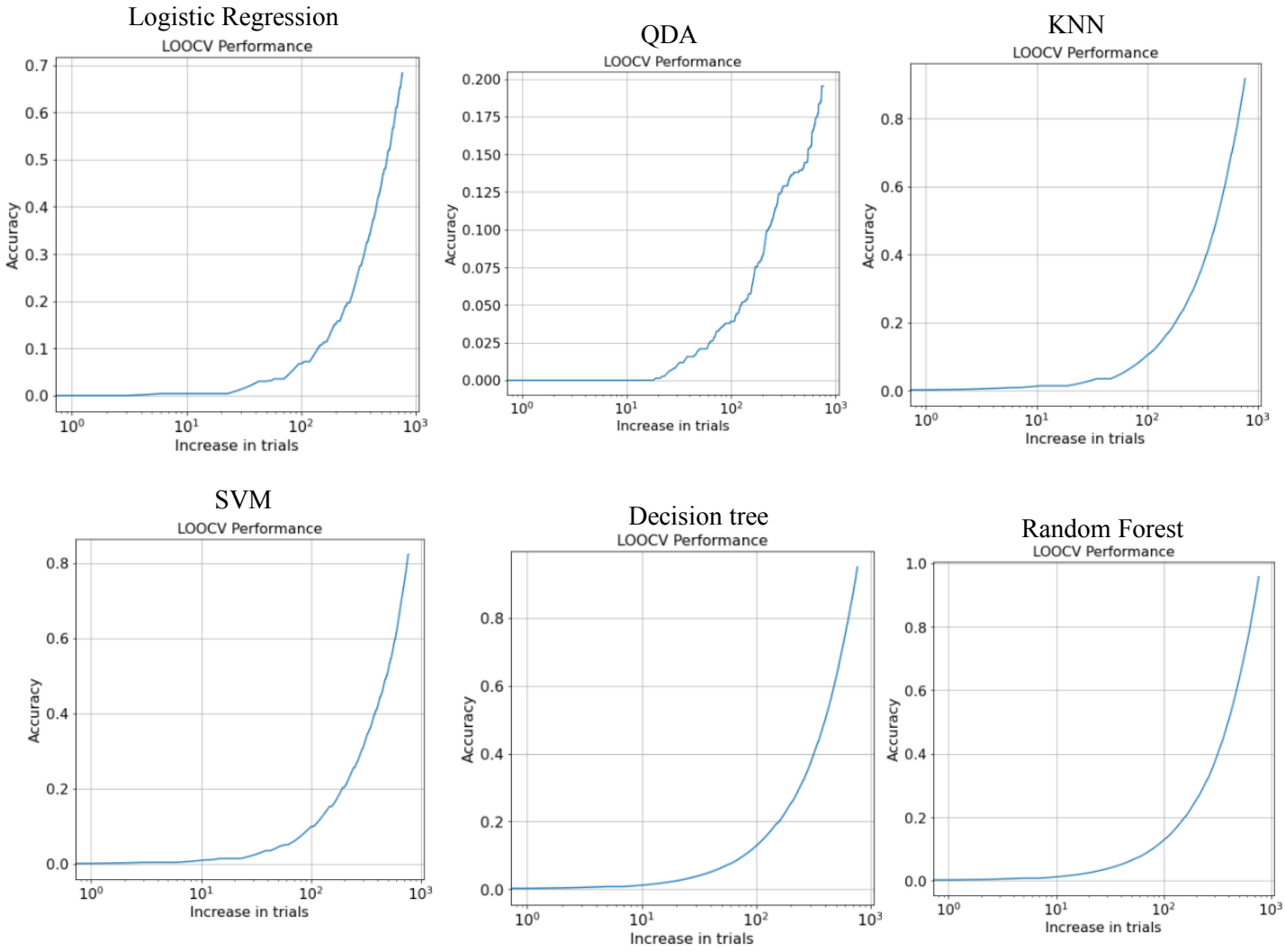

Figure 20: LOOCV performance

7.2. Feature Reduction

As explained in the feature reduction section that the correlation between Orientation with other features is zero or in other way the Orientation data column is uncorrelated with the input features the output. Hence the uncorrelated data is dropped from the input feature and model is trained again to observe the impact of uncorrelated data on the model. It was strange to observe that the model performance has increased by dropping one of the columns but it is also expected as well because if case of logistic regression one feature is added which having no impact on the model performance that will work as the penalty parameter in training as it will add error while training the model.

Based on above experience of removing the uncorrelated data while training the model, triggers another thing that if we remove more features based on the variance ratio obtained from the PCA than the performance of the model will further enhanced. Hence PCA is used with number of components as 6 to remove one feature and to convert correlated data to uncorrelated.

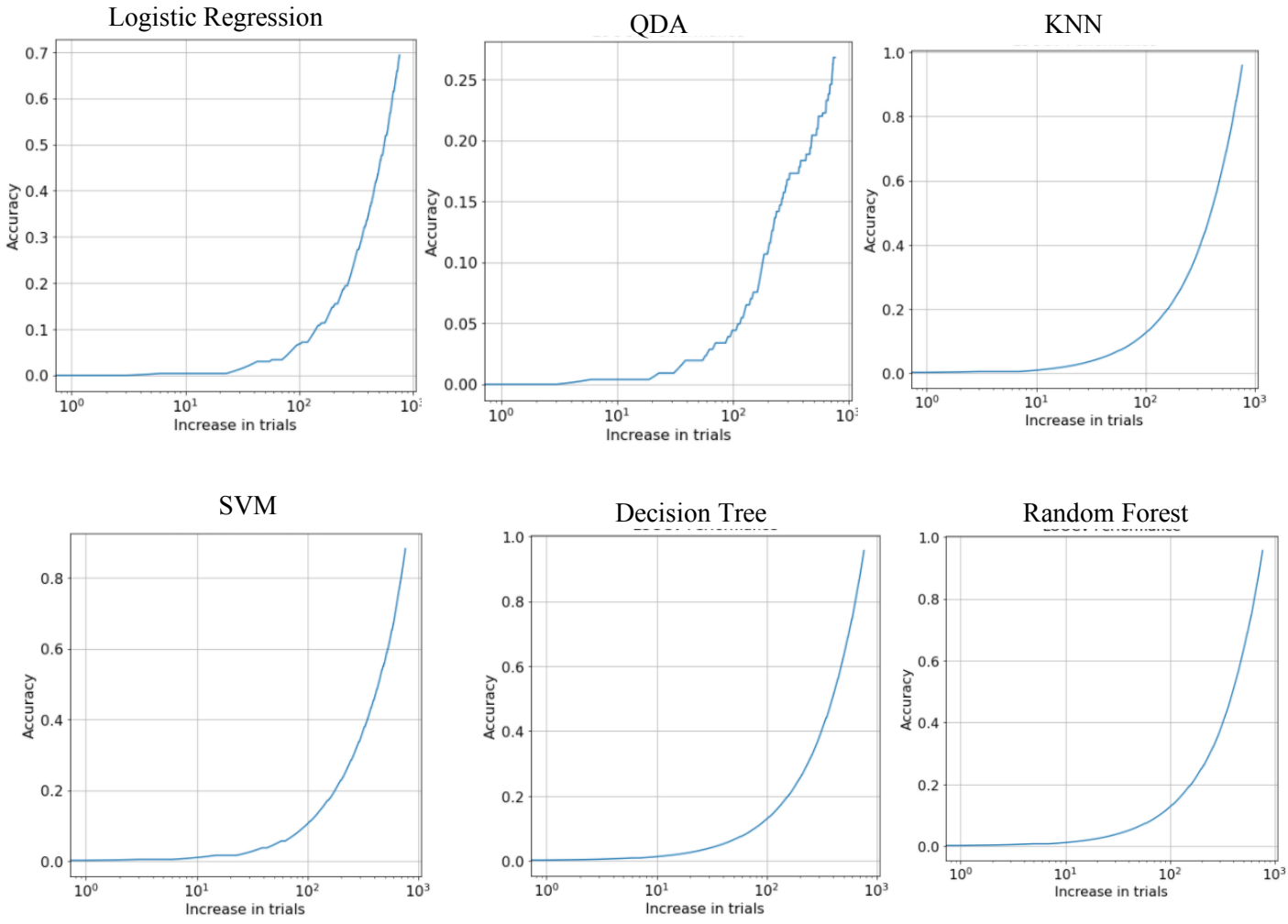

Figure 21: LOOCV Performance after Feature Reduction

As shown in Fig.21, PCA is used to reduce the input features and LOOCV is used as cross validation technique to train the above listed models for classification. For few models the performance has increased namely KNN, SVM, Decision Tree and Random Forest. Another thing observed here is that if we reduce one feature or we use PCA by reducing one feature, we get almost same performance of the model.

8. Performance

To classify heating load using seven features of the building, six models are designed with Linear namely Logistic Regression and Non-linear decision boundary namely QDA, KNN, SVM, Decision Tree and Random Forest. Each model is again trained using LOOCV cross validation technique to generalize the model parameters and improve the performance of the model. The models are also trained with feature reduction technique by dropping one of the features namely Orientation and performance was greatly enhanced. By applying PCA with LOOCV model is again trained by reducing the number of features and by converting correlated data to uncorrelated data as well as using LOOCV for improving the performance of the model. And lastly, the models are trained with only feature reduction technique using PCA without LOOCV and performance of the model is noted.

To summaries all the performance, Fig. 22 is represented as shown which provides bar graph of all model performance for test data. As shown, the Logistic Regression and Quadratic Discriminant Analysis provides worst result compare to other models, while KNN, SVM, Decision Tree and Random Forest provides better results. Using only PCA for feature reduction has low impact as compared to feature reduction by dropping one feature directly. As well as there is no much difference in the output for PCA with LOOCV and PCA with LOOCV technique. Decision Tree and Random Forest are having almost equal performance but Random Forest is chosen as in Random Forest the number of features are chosen randomly and performs better in test data.

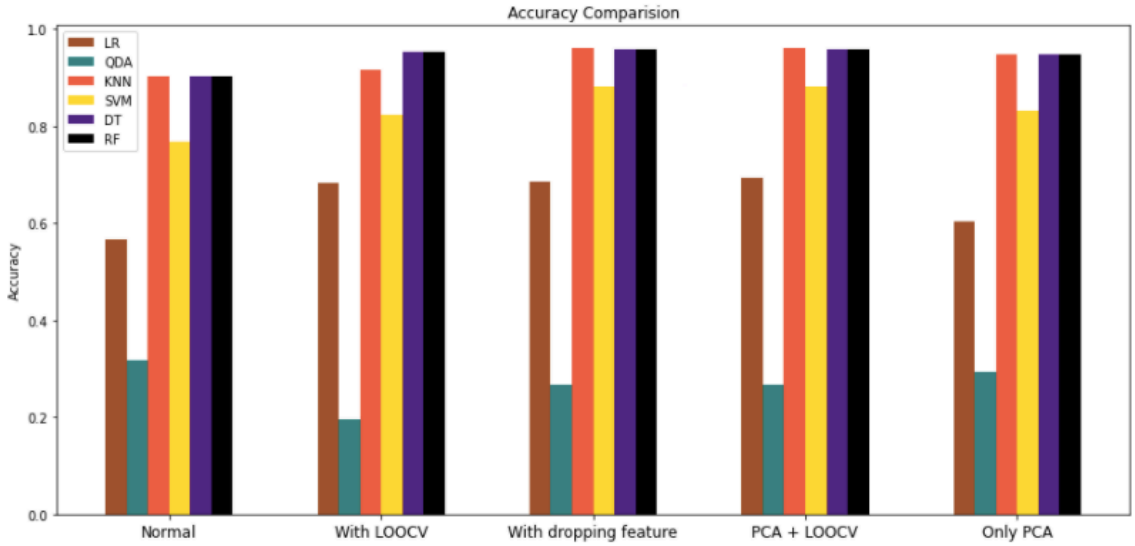

Figure 21: Performance Comparison of all Models

By choosing the model as Random Forest, the performance comparison is again done to understand small difference of the accuracy as shown in Fig. 22. As shown, the performance of the model is highest with LOOCV without feature reduction which achieves almost **95 % as the test accuracy.** With using LOOCV another advantage is to obtain the more generalize model as compared to overestimate model using validation set approach.

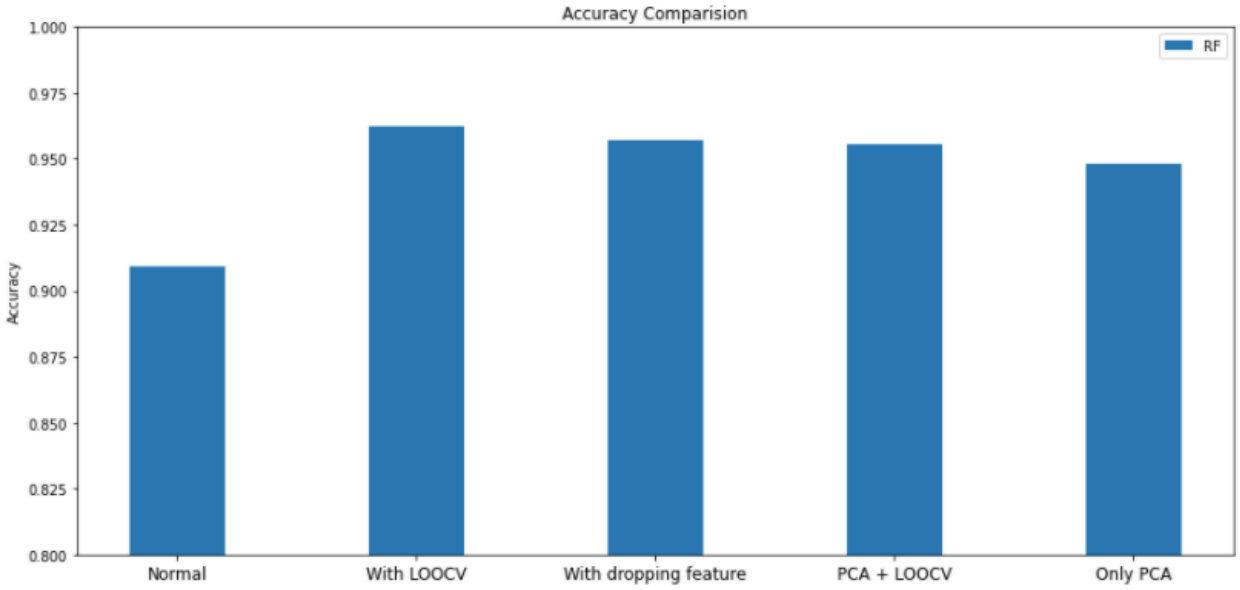

Figure 22: Performance of the Random Forest

Confusion matrix is developed to understand mis-classification of each class and to find other performance parameters of the model. As shown in Fig. 19, the confusion matrix represents Actual and predicted labels from class 0 to 9. The performance parameters are found using following equations:

$$Accuacy = \frac{TP + TN}{Total\ number\ of\ data\ points} * 100$$

$$Precision = \frac{TP}{Predicted\ Positive} = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

Where *TP* is true positive, *TN* is true negative, *FP* is false positive while *FN* is false negative.

To understand the performance of precision recall for each class, Fig. 23 is plotted combining precision and recall for all classes. As represented the precision-recall curve for class 1 is the highest which states that class 1 was classified correctly with very less false positive and false negative values. For class 9 the performance is worst as compared to other classes and that could

be understandable as the dataset is not uniformly distributed as well as each class were defined based on the approximation to the nearest values.

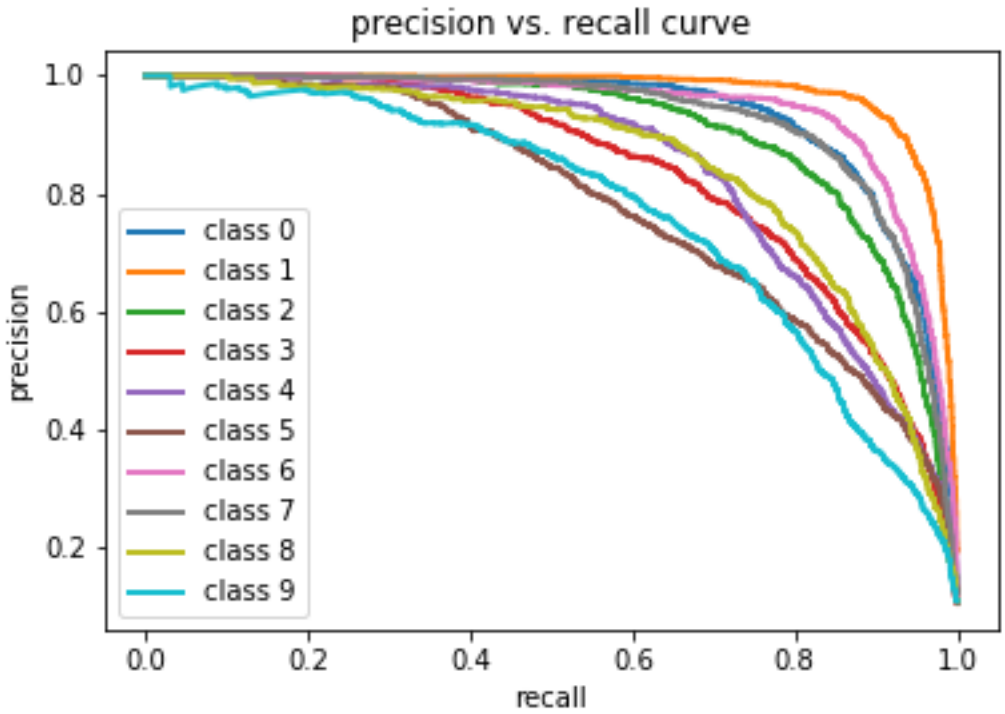

Figure 23: Precision Recall Curve

Furthermore, to understand the overall performance of the model, ROC (Receiver operating characteristic) curve is plotted by taking average of all the class performance as shown in Fig. 24. In figure, dotted line represents the worst model prediction with 50% probability of getting success.

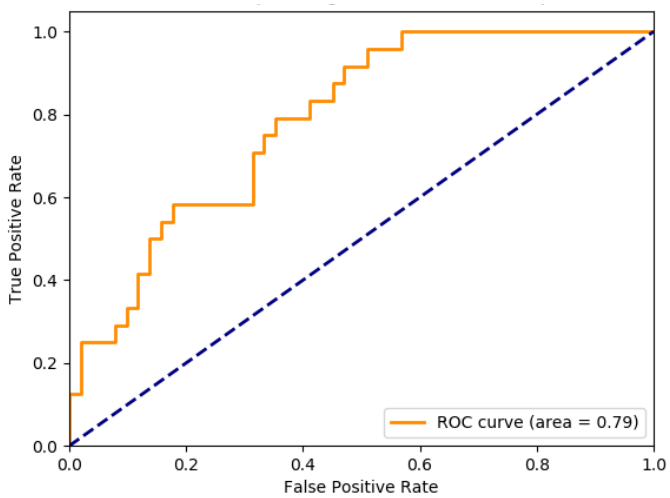

Figure 24: ROC curve

The area under the curve represents the possibility of getting True positive more and improving the accuracy. The area under the curve obtained for Random Forest model with average of all the classes is 0.79.

9. Conclusion

The aim of this project is to classify Heating Loads in the building to determine the specifications of the heating equipment needed to maintain comfortable indoor air conditions. The dataset consists of 8 features while only heating load is considered as labels in this project with small data points to train and test the model performance. Firstly, the heating load is available with floating values and to convert into classification problem, the labels are divided into classes by defining each class with range of values of the labels. The dataset is not uniformly distributed with the number of classes defined and hence desired performance may vary.

In data cleaning, the statistics of the model is found and observed that the distribution of the dataset does not consists of any missing values. While plotting the data in box plot, it was observed that there were no outliers in the dataset. Hence the dataset does not require any cleaning methods to perform. While plotting the correlation of the dataset it was observed that Orientation column is not correlated with the any other feature. While the Glazing distribution and Glazing distribution area are having very less correlation, which is also understood as if the area of the distribution is high does not mean that the heating load required would be high. While the relative compactness and the surface area are highly correlated but inversely. Based on the correlation matrix, the dataset is greatly understood and by plotting in heatmap, the visualization of the correlation is greatly enhanced.

Baseline model is developed with six models namely Logistic Regression, QDA, KNN, Decision Tree, SVM and Random Forest. Initially all the models provide bad result as the validation set approach is used and hence LOOCV cross validation technique are used to generalize the model performance. Logistic Regression and QDA still does not perform well as the dataset are highly non-linear and the decision boundary where linear and quadratic in nature for the respective models. In feature reduction, two technique is used for model training namely by dropping Orientation column from training the model as it consists of zero correlation among all other input features. Another technique used for feature reduction is by using PCA with number of components as six which means by dropping one feature and converting correlated data to uncorrelated. With all the feature reduction technique all six models were trained and their performance is noted. Lastly, PCA input features are used to train the model with LOOCV cross validation technique which helps to further enhance the performance of the model.

In models with non-linear decision boundary, **Random Forest** performs best with **95% test accuracy** by using PCA with LOOCV technique. As the model is generalized with the test data by using cross validation technique and by using PCA correlation data converted to uncorrelated as well as the number of features required to train the model has reduced. Hence the model is neither overfitted nor under fitter and the best or the optimum performance model is obtained by using Random Forest. To verify and understand the performance of the model, graphs are plotted for precision and recall for each class as well as RoC curve is plotted by averaging the output of each class and the area under the curve obtained is **0.79.**

In conclusion, with the help of machine learning the energy efficiency increase as we obtain some inference regarding the type of input feature effected on the output and some correlation of the input features. By using the model the heating load of the building will be obtained and will help

to improve the performance of the equipment by proper design the product. Similar to the heating load, prediction or classification of the cooling load may be possible with the given dataset and will be considered as the future scope of the project.

10. References

1. https://archive.ics.uci.edu/ml/datasets/Energy+efficiency
2. A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012
3. https://scikit-learn.org/stable/
4. https://colab.research.google.com/github/shranith/Colab-intro/blob/master/Colab_intro.ipynb
5. https://seaborn.pydata.org/
6. https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Energy%20Efficiency
7. https://rstudio-pubs-static.s3.amazonaws.com/244473_5d13955ea0fd4e5e9d376161b956e9dc.html
8. European Commission, Directive 2002/91/EC of the European Parliament and of the Council of 16th December 2002 on the energy performance of buildings, Official journal of the European Communities, L1/65–L1/71, 04/01/2003.
9. W.G. Cai, Y. Wu, Y. Zhong, H. Ren, China building energy consumption: situation, challenges and corresponding measures, Energy Policy 37 (6) (2009) 2054–2059.
10. Z. Yu, F. Haghigrat, B.C.M. Fung, H. Yoshimo, A decision tree method for building energy demand modeling, Energy and Buildings 42 (2010) 1637–1646.