Assessment of heating and cooling load requirements of building using machine learning techniques.

GROUP NO. :- 20

ABHIJEET SARKAR(193100068)

JAINIL SHAH(193100071)

SAURABH MANDAOKAR(193100081)

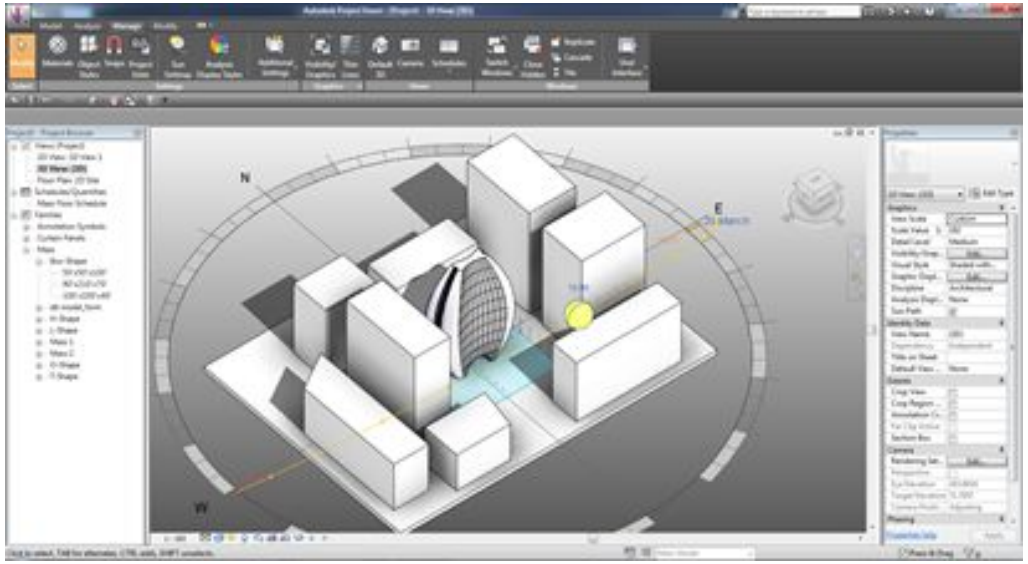
How and why a building consumes energy ?

If glazing area of building is very high this may reduce the need of illumination till dusk time but it may increase the need to cool the inside of a building due to high exposure towards sunlight.

This also depends on what conditions are suitable for human comfort.

How data was obtained ?

- 18 cubes of dimensions (3.5 * 3.5 * 3.5) taken to make 12 different forms of buildings.
- All buildings simulation was done in Ecotect software and each building has a volume of 771.75 cubic meter.



Description of data features

features:

- X1: Relative Compactness
- X2: Surface Area
- X3: Wall Area
- X4: Roof Area
- X5: Overall Height
- X6: Orientation
- X7: Glazing area
- X8: Glazing area distribution

outputs:

- Y1: Heating Load
- Y2: Cooling Load

Operations performed on original data sets

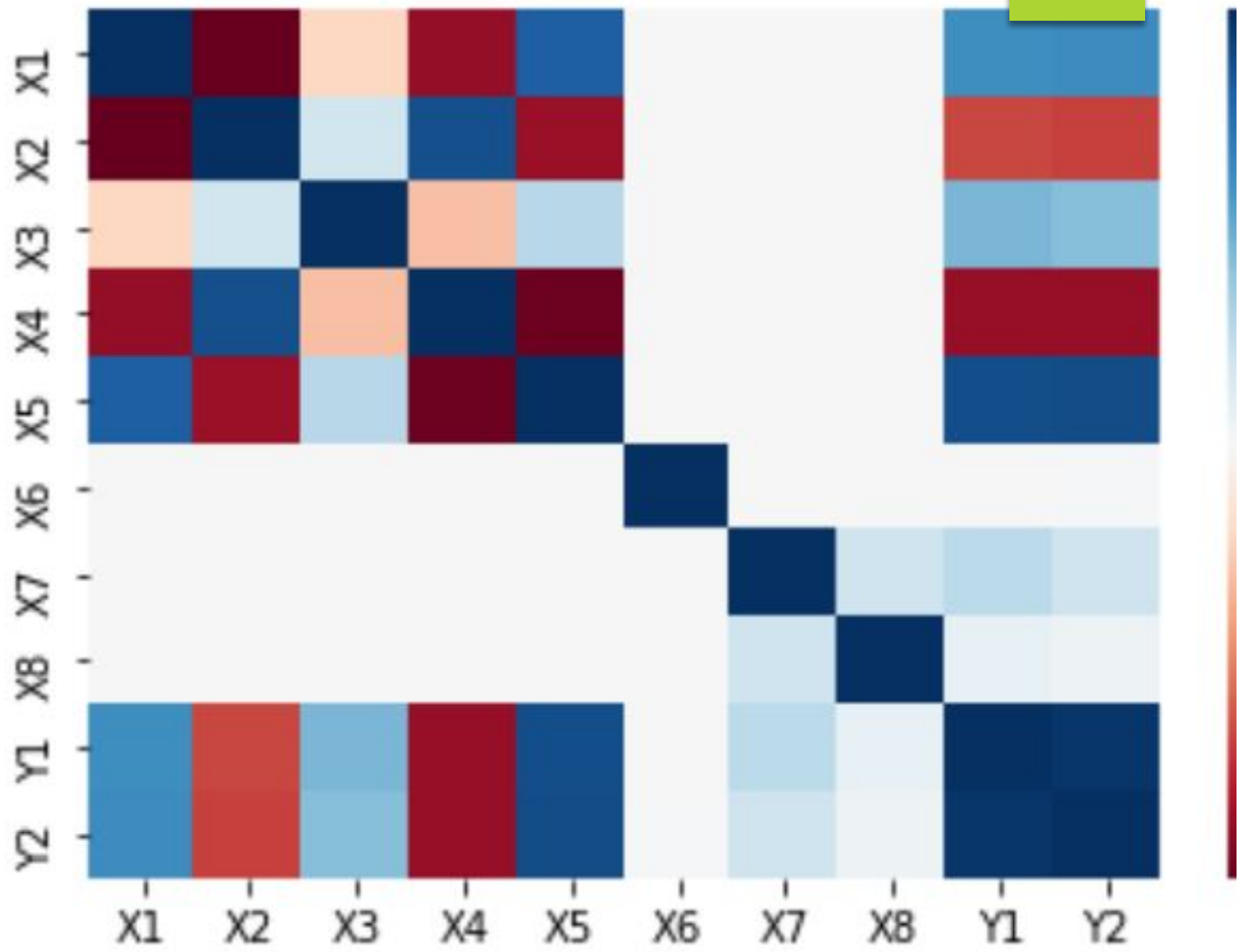
10% of data was knocked from each column, which resulted in 77 empty cells in each column of the features.

Imputation of data into these vacant cells were done using various methods and prediction models.

Perfect model for the imputed data was found that can predict well.

Heatmap

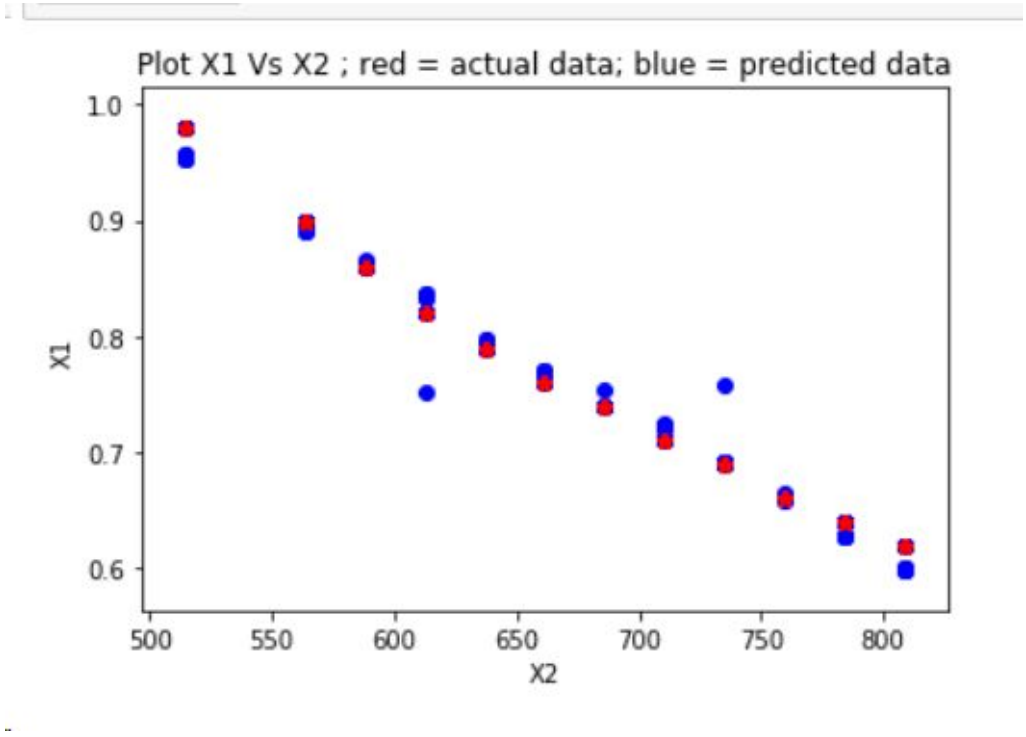
visual analysis of multi collinearity between predictors.



Imputation of X1 (relative compactness)

- X1 and X2 show a linear behavior with each other. By feature selection it was found that X1 predictions are best when we use features X2 as well as Y1.

1) By multiple linear regression
2) Decision Tree
3) KNN(k=5)
4) Random Forest

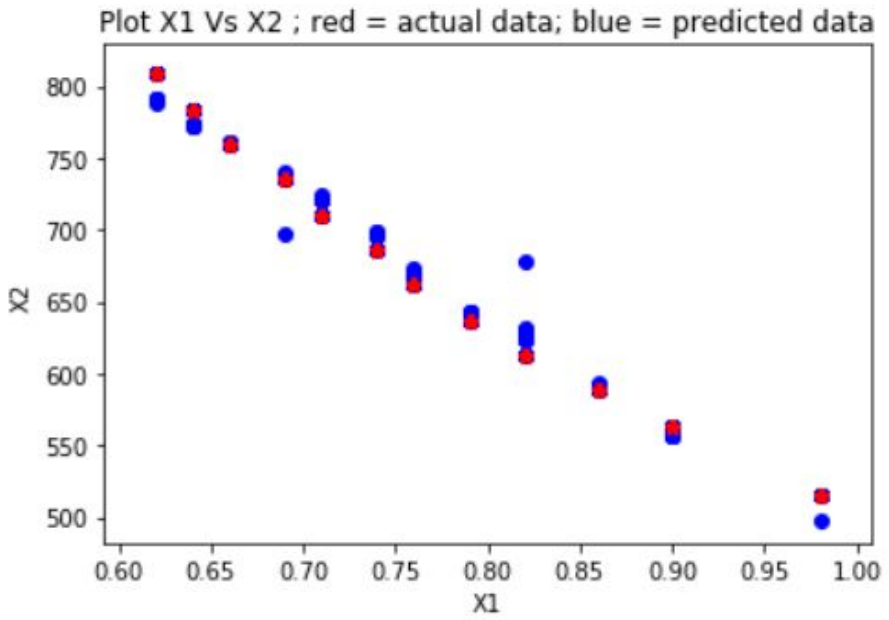


Method	Multiple Linear Regression	Decision Trees Regressor	KNN(K = 5)	Random Forest
Accuracy	0.997561	0.9989629 2	0.99896	0.99916

Imputation of X2 (Surface Area)

- By using modified X1 values predicted X2 values by following method using X1 and Y1

1) Multiple linear regression:
2) Decision Tree
3) KNN(k=5)
4) Random Forest



Method	Multiple Linear Regression	Decision Tree Regressor	KNN (K = 5)	Random Forest
Accuracy	0.9978	0.9991	0.98791	0.99919

Imputation of X5 (overall height)

- X5 takes only 2 values 7 and 3.5 as overall heights.
- These 2 values were encoded using **label encoder**
- Therefore the predictions for X5 can be expected to be fairly accurate using X1, X2, X4, Y1 and Y2

Methods used:

- KNN regressor : accuracy = 0.998
- Logistic regression : accuracy = 0.993

```
[[384   0]
 [  1 383]]
```

Confusion matrix for KNN

```
[[384   0]
 [  5 379]]
```

Confusion matrix for Logistic regression

Imputation of X4 (roof area)

- the features selected for predictions of X4 are X1,X2,X5,Y1 and Y2.
- First, we tried to fit a KNN regression model. The prediction values accuracies were obtained by comparing with original data which came out to be
- 99.99 % accurate at K = 5.

Imputation of X3 (Wall Area)

Imputation of X3 includes features X1,X2,X4,X5,Y1 and Y2.

KNN and decision trees produced very accurate predictions closer to original data.

Imputation of X7 (glazing area percentage)

Steps:

- X7 has values as a percentage of floor area. The prediction can be done by
- Converting the percentage values to numerical values by using floor area data from X4 values. Multi linear regression and KNN regression were used for prediction on this data.
- Encoding the data using Level encoder and performed KNN classifier for predictions using X1, X2, X3, X4, X5, Y1 and Y2

X7 by MLR:

Accuracy = 0.94

X7 by KNN:

Prediction accuracy = 0.99

```
[[ 48   0   0   0]
 [  0 240   0   0]
 [  0   0 240   0]
 [  0   0   0 240]]
```

Confusion matrix for KNN classifier predictions with original data

Imputation of X6

(Orientation)

- As X6 as no correlation with any other features, so imputation was done using Y1 and Y2 only.
- Also the data was class type.

Method to predict X6	Accuracy when Compared to Original X6	Model accuracy (of fitting data)
LDA	0.923	0.266
QDA	0.923	0.2708
ANN	0.923	0.3346
Decision trees	0.925	0.991

Imputation for X8
(Glazing area distribution)

- Like X6, X8 is also not correlate with any other feature so it is imputed using Y1 and Y2.
- Also X8 is class type data

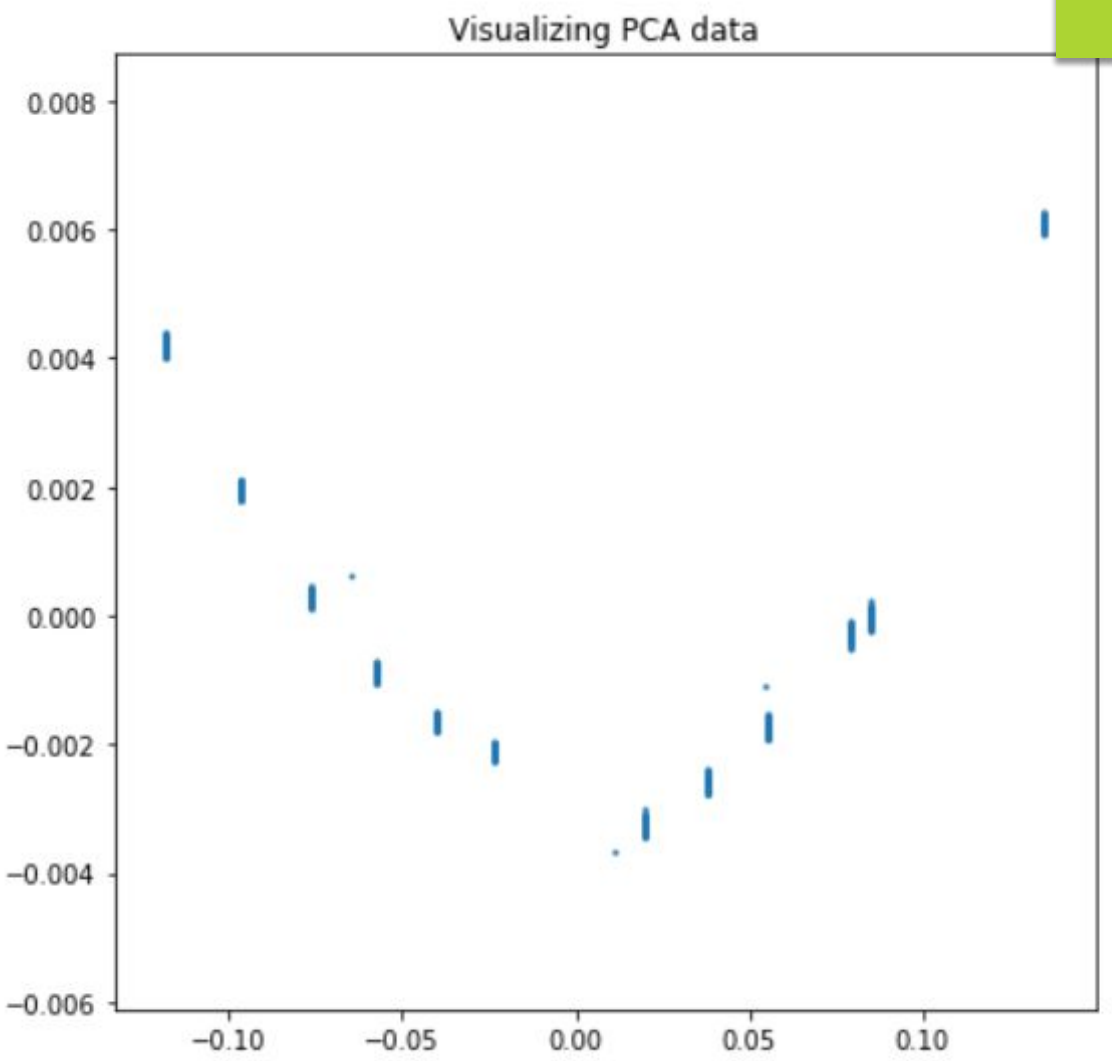
Method to predict X8	Accuracy when Compared to Original X8	Model accuracy (of fitting data)
LDA	0.91406	0.23733
QDA	0.91796	0.24023
KNN with K = 10	0.92708	0.35021
ANN	0.91666	0.24023
Decision trees	0.92778	0.99276

Model generation process

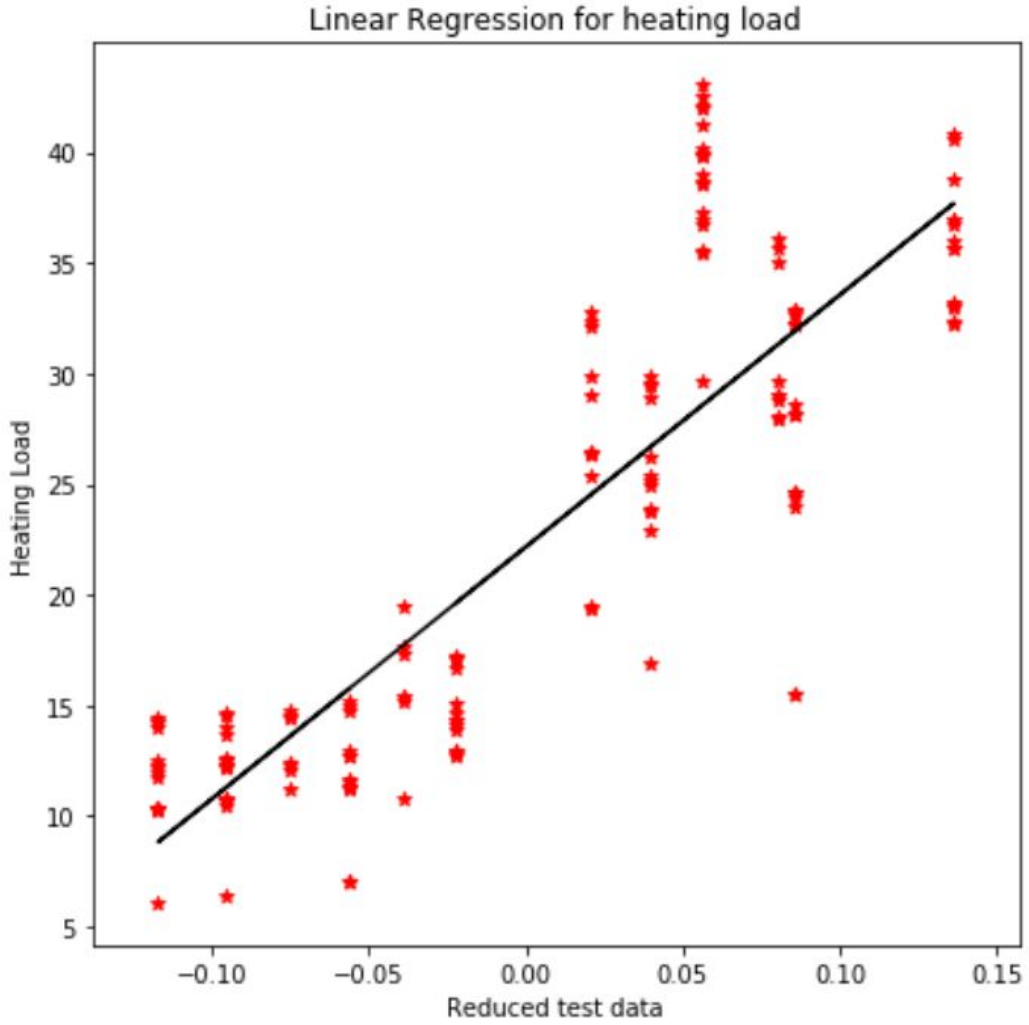
Reducing	Reducing the data set of features into a single dimension for PCA analysis.
Trying	Trying various models on this reduced data set.
Selecting	Selecting appropriate models to work on with imputed data set.
Trying	Trying to fit selected models on the imputed data sets and selecting the best model.

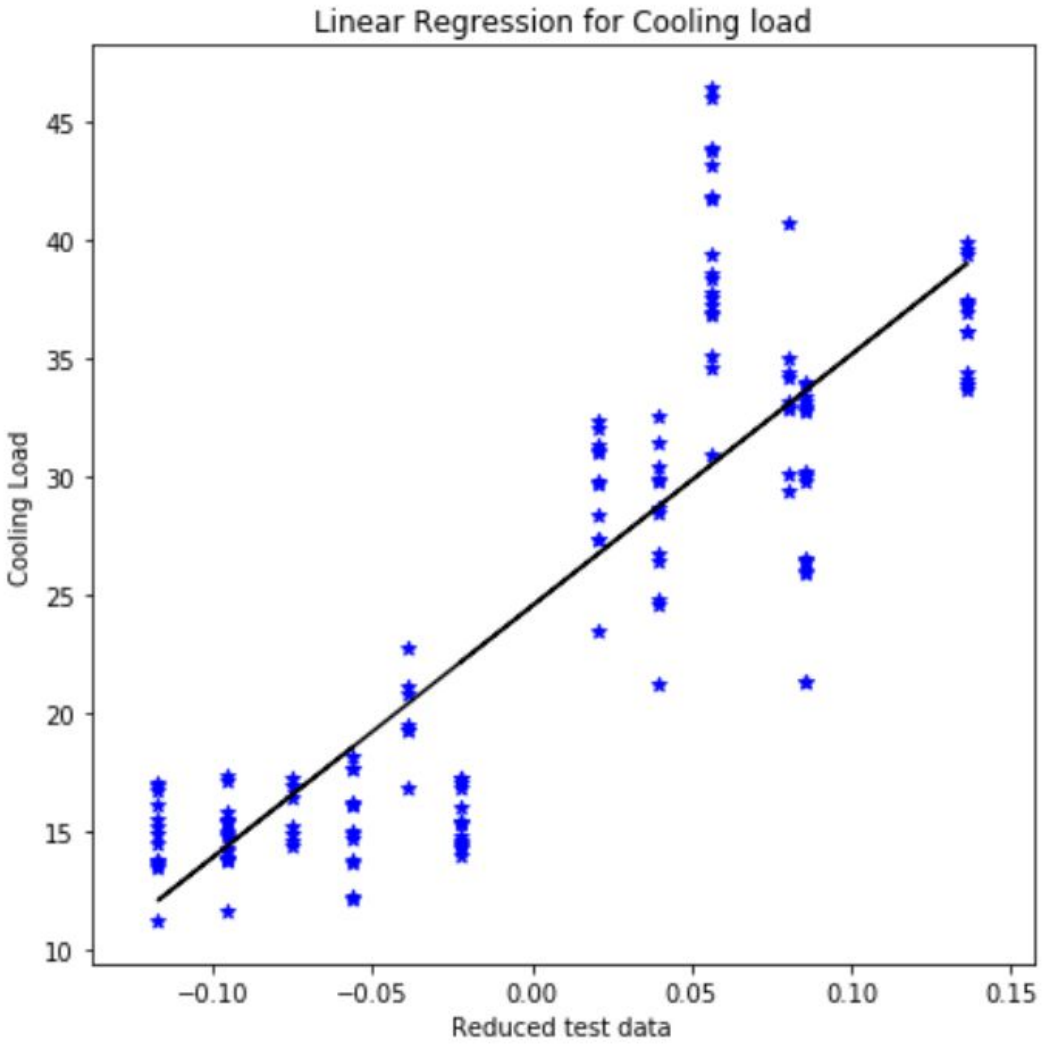
Reduced data set

- Visualization of reduced data set. Data scattered in 12 clusters.



Linear regression on reduced data set





Model summary for Heating Load

OLS Regression Results

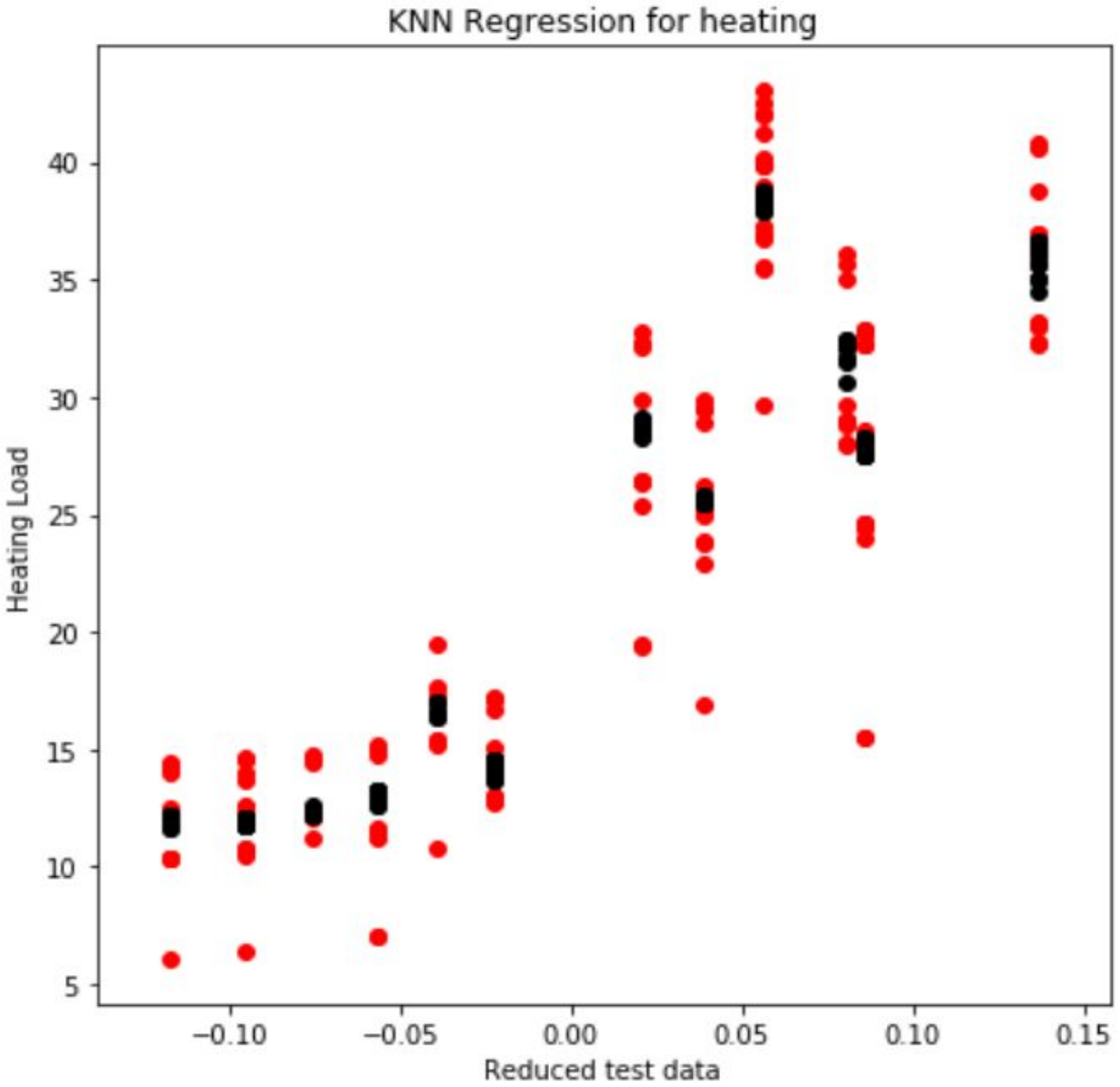
Dep. Variable:	y	R-squared (uncentered):	0.128
Model:	OLS	Adj. R-squared (uncentered):	0.126
Method:	Least Squares	F-statistic:	89.64
Date:	Fri, 25 Oct 2019	Prob (F-statistic):	6.09e-20
Time:	09:31:14	Log-Likelihood:	-2788.7
No. Observations:	614	AIC:	5579.
Df Residuals:	613	BIC:	5584.
Df Model:	1		
Covariance Type:	nonrobust		

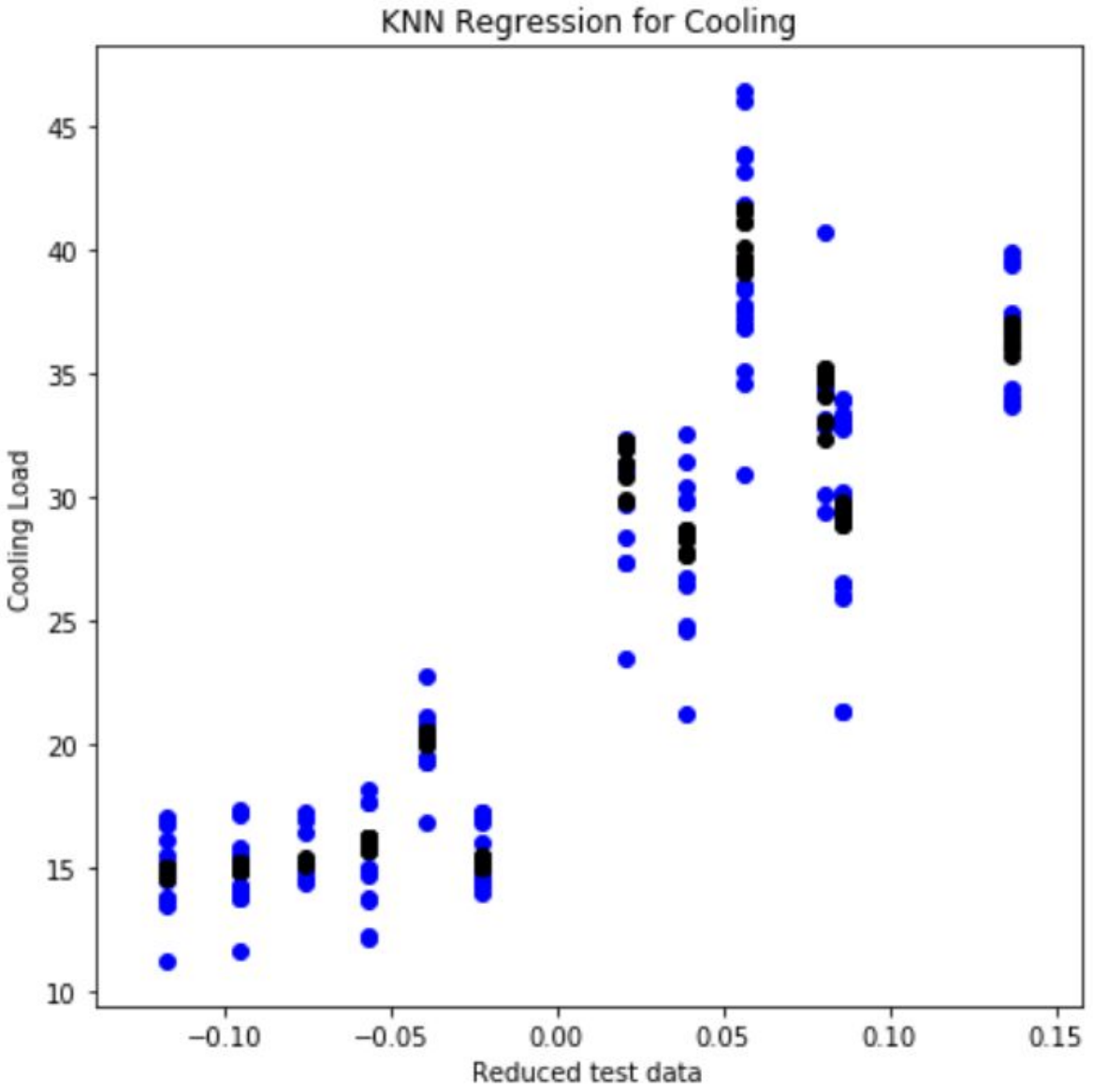
Model summary for cooling Load

OLS Regression Results

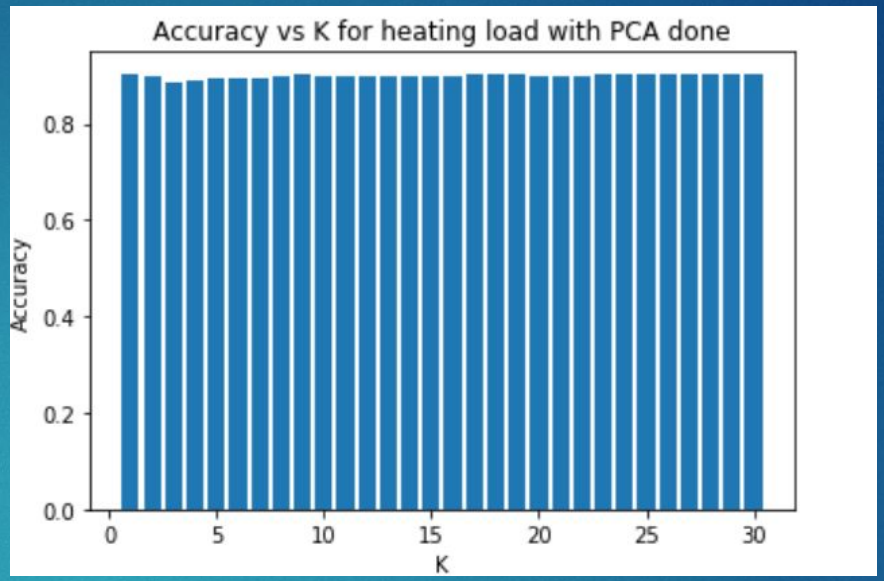
Dep. Variable:	y	R-squared (uncentered):	0.095
Model:	OLS	Adj. R-squared (uncentered):	0.094
Method:	Least Squares	F-statistic:	64.39
Date:	Fri, 25 Oct 2019	Prob (F-statistic):	5.22e-15
Time:	09:31:14	Log-Likelihood:	-2847.6
No. Observations:	614	AIC:	5697.
Df Residuals:	613	BIC:	5702.
Df Model:	1		
Covariance Type:	nonrobust		

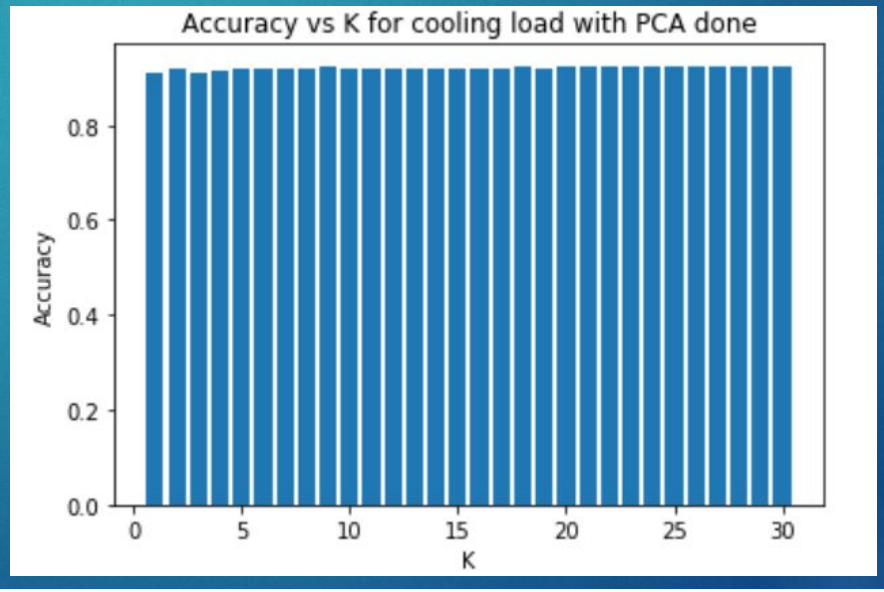
KNN regression on reduced dataset



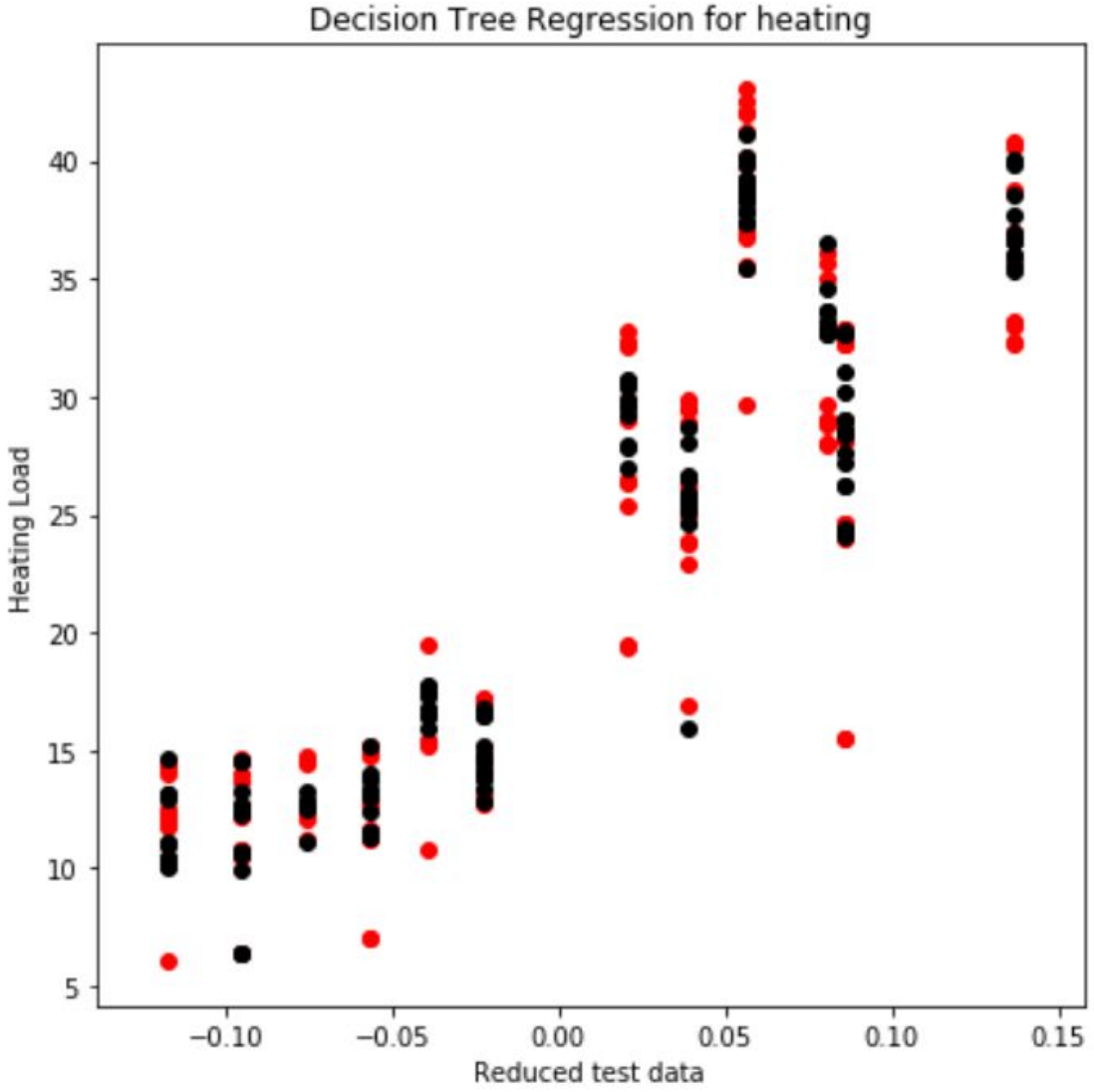


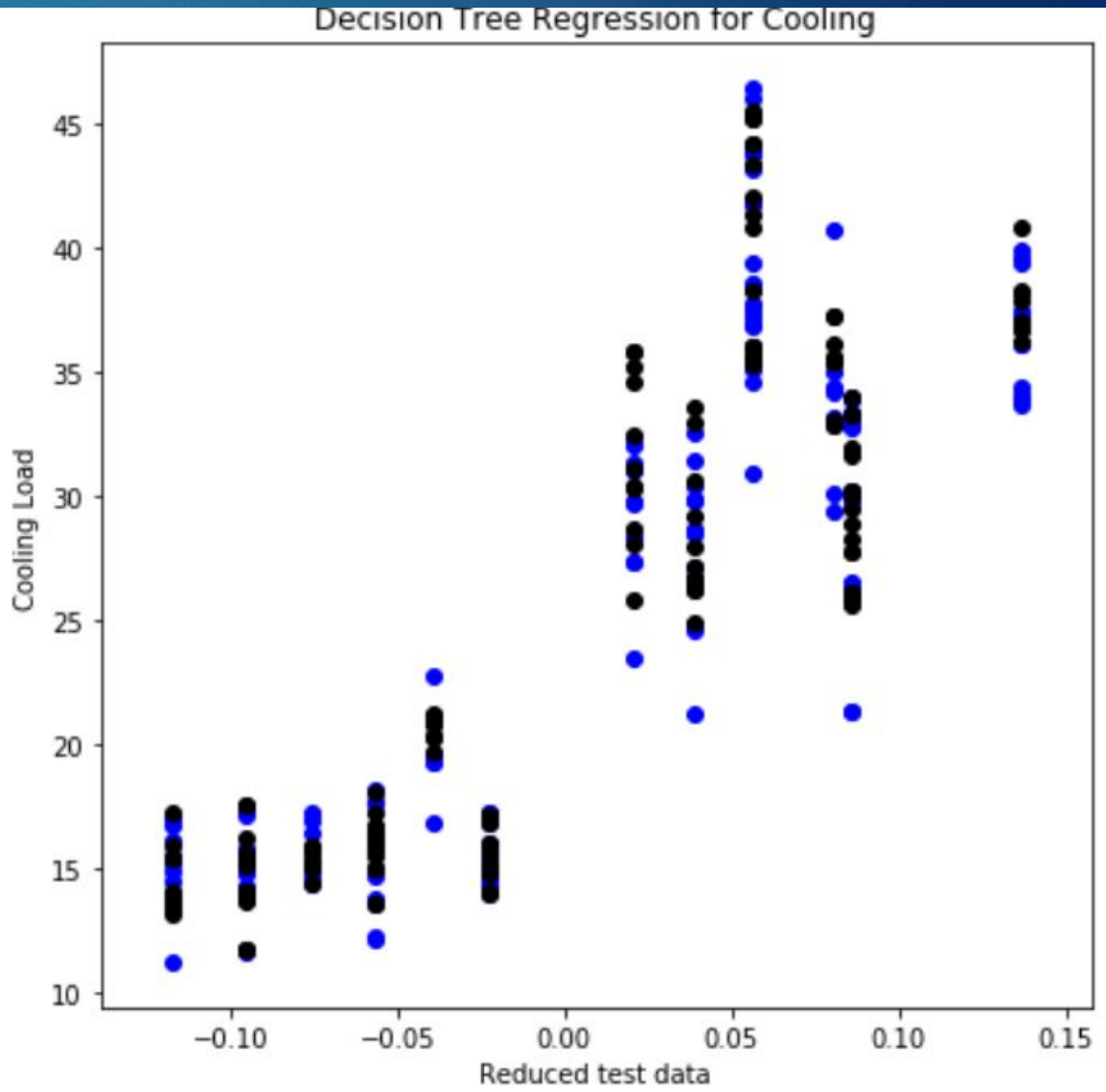
- model accuracy in case of KNN :
- 1. for heating load = 0.9043
- 2. for cooling load = 0.9250



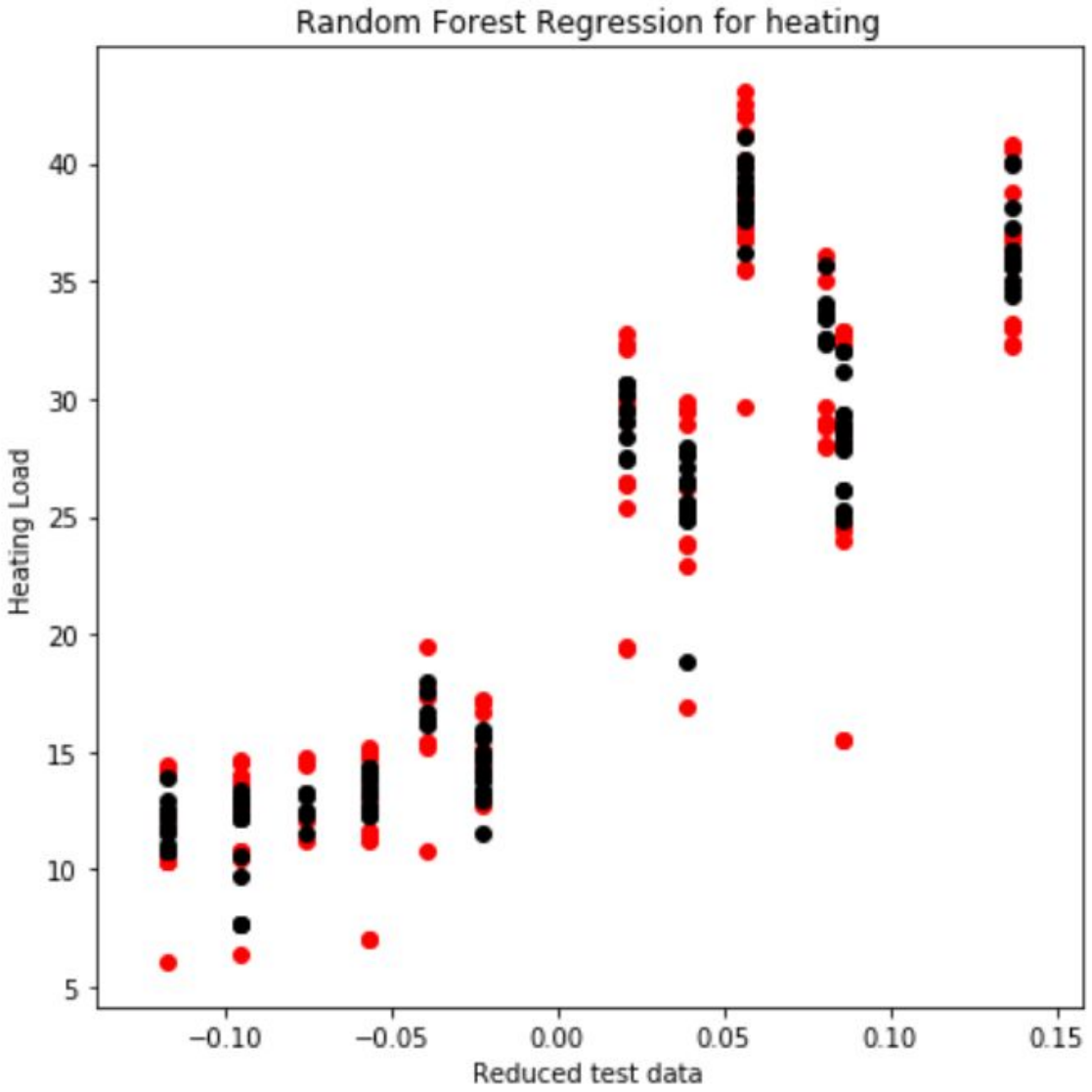


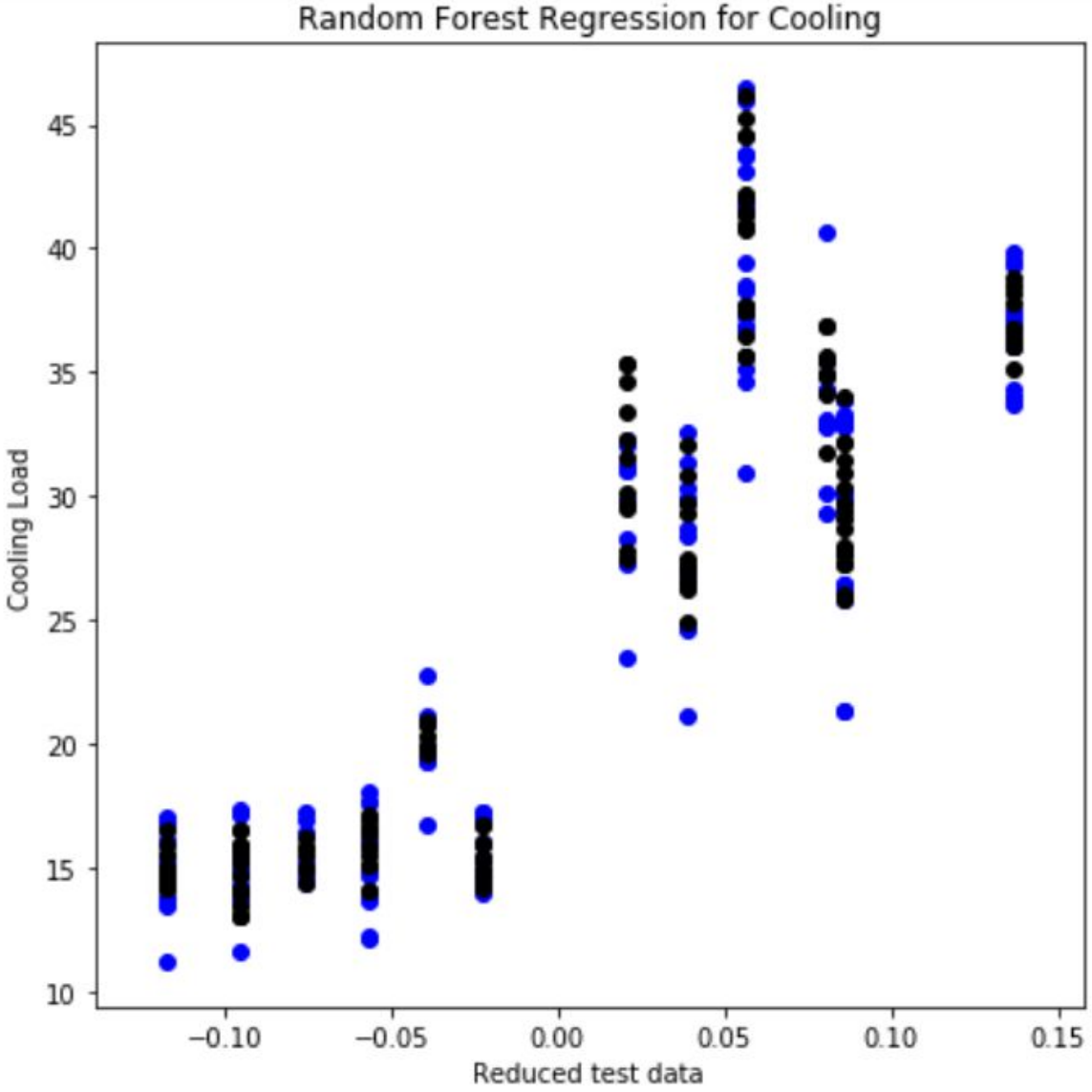
Decision trees on reduced dataset





Random forests on reduced datasets



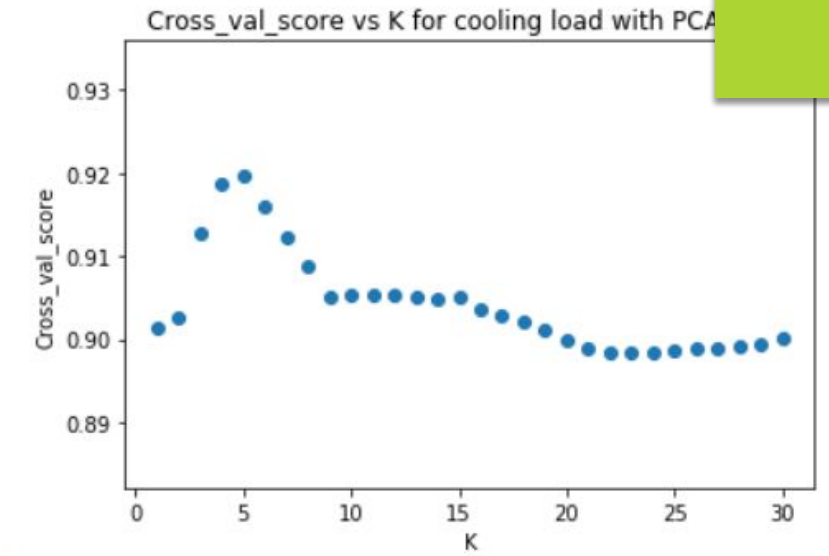


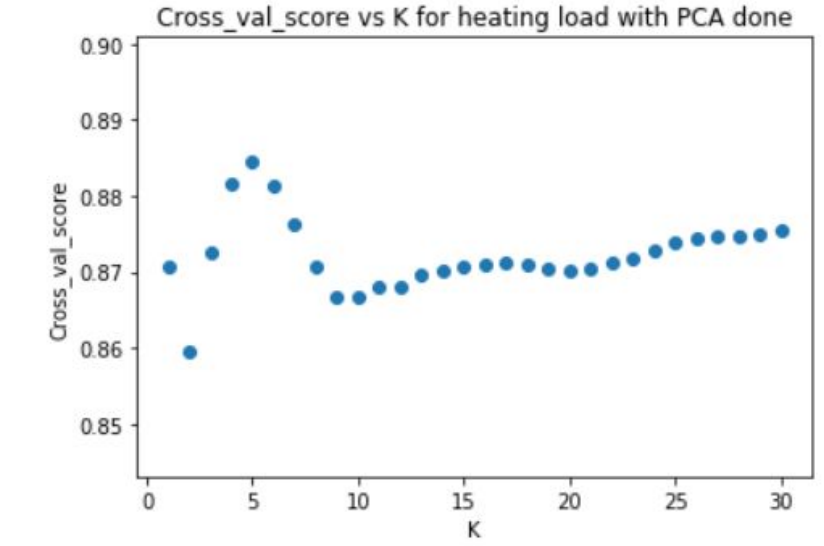
Summary for reduced data

PREDICTION MODEL	OUTPUT	R2_SCORE
linear regression	heating load	0.7268
	cooling load	0.729
KNN regression	heating load	0.897
	cooling load	0.921
Decision trees	heating load	0.872
	cooling load	0.907
random forests	heating load	0.885
	cooling load	0.913

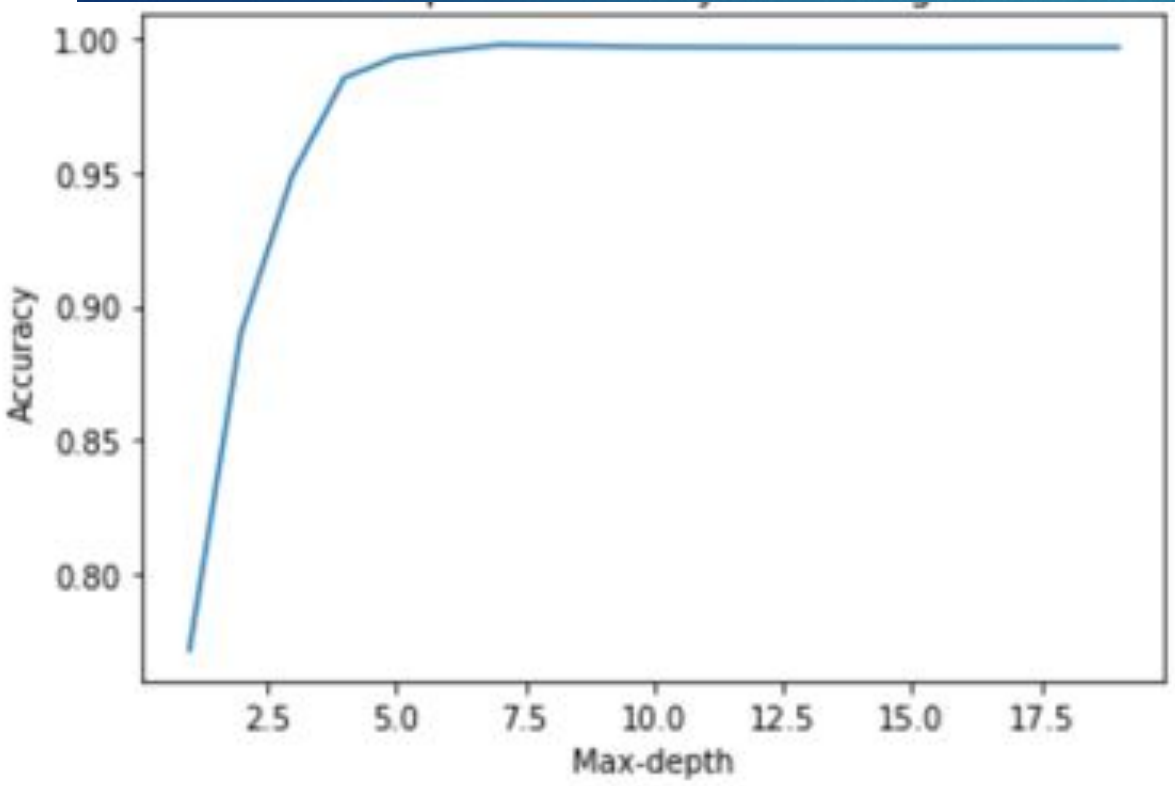
fitting knn regression on imputed data

- Thus maximum Cross_val_score = 0.88 is for k = 5 for heating load.
- Thus maximum Cross_val_score = 0.91 is for k = 5 for cooling load.

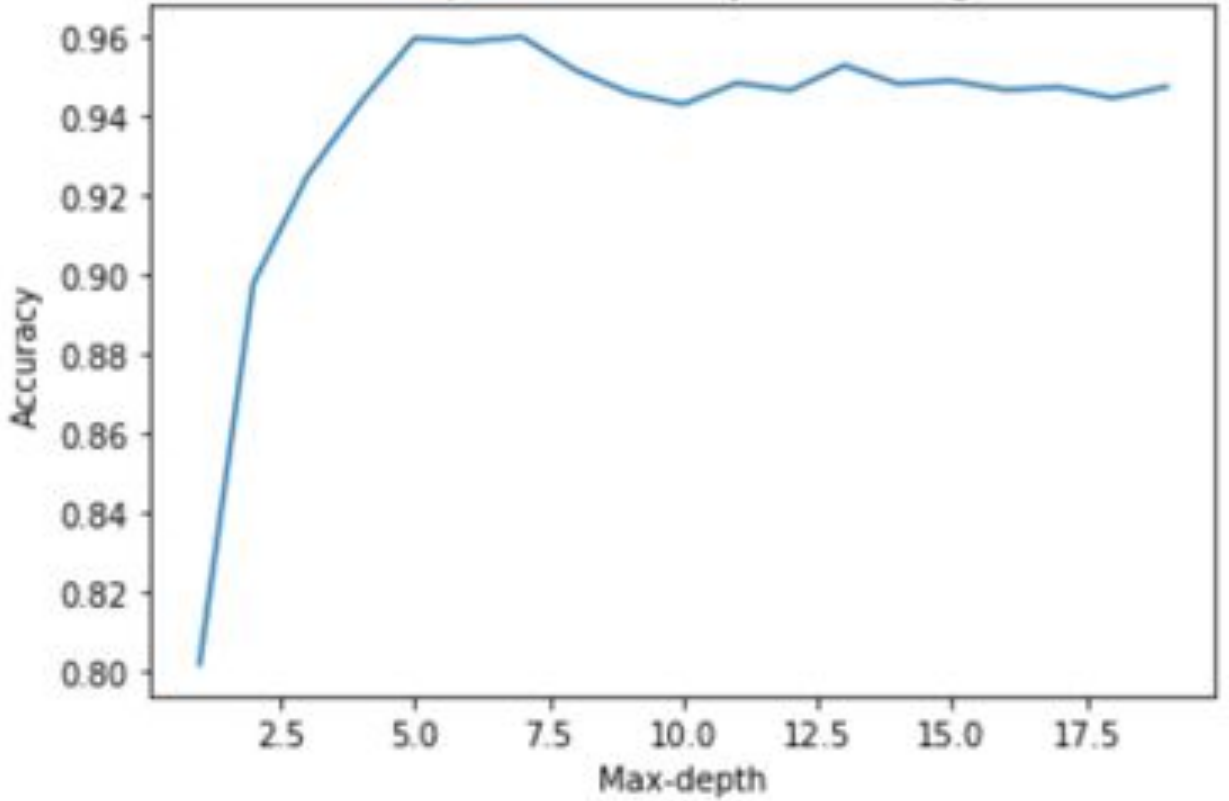




Fitting decision trees on imputed data

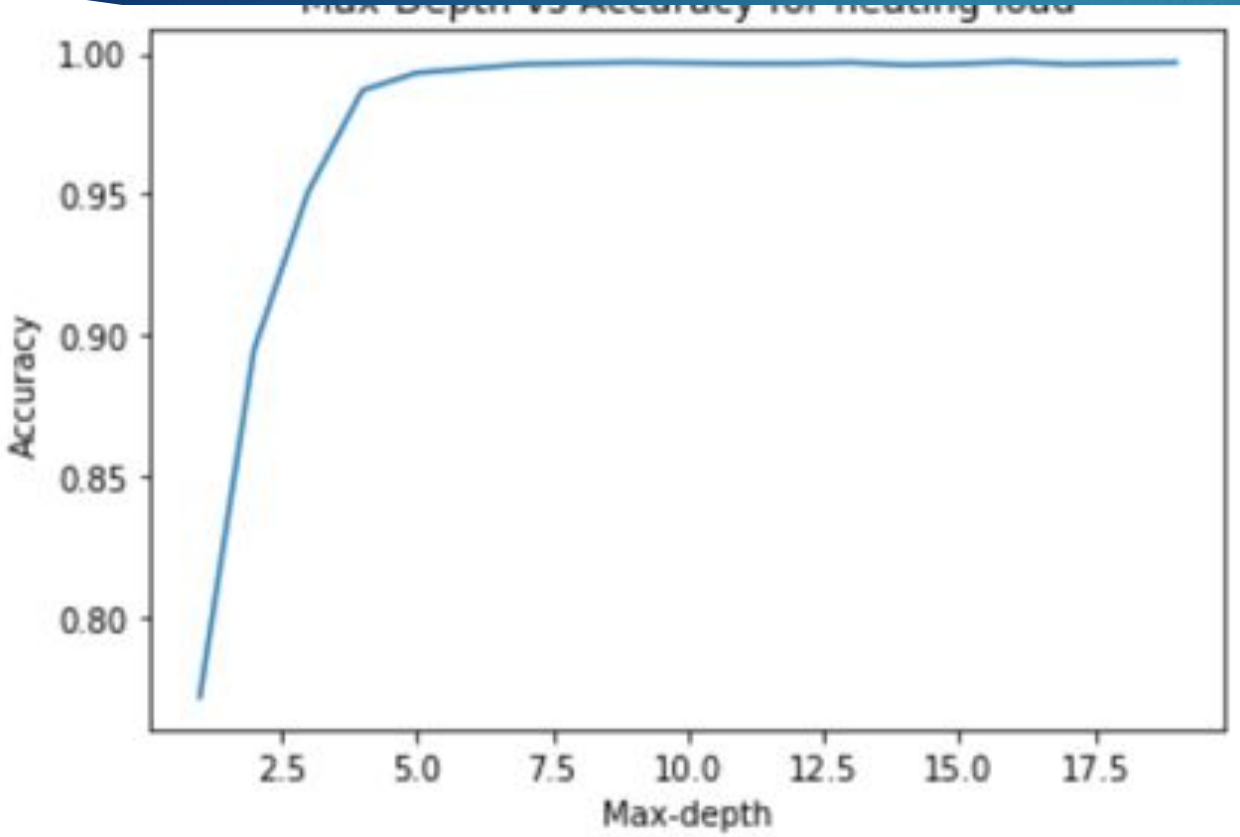


Max depth vs accuracy for heating load
Max depth = 7 accuracy = 0.9973

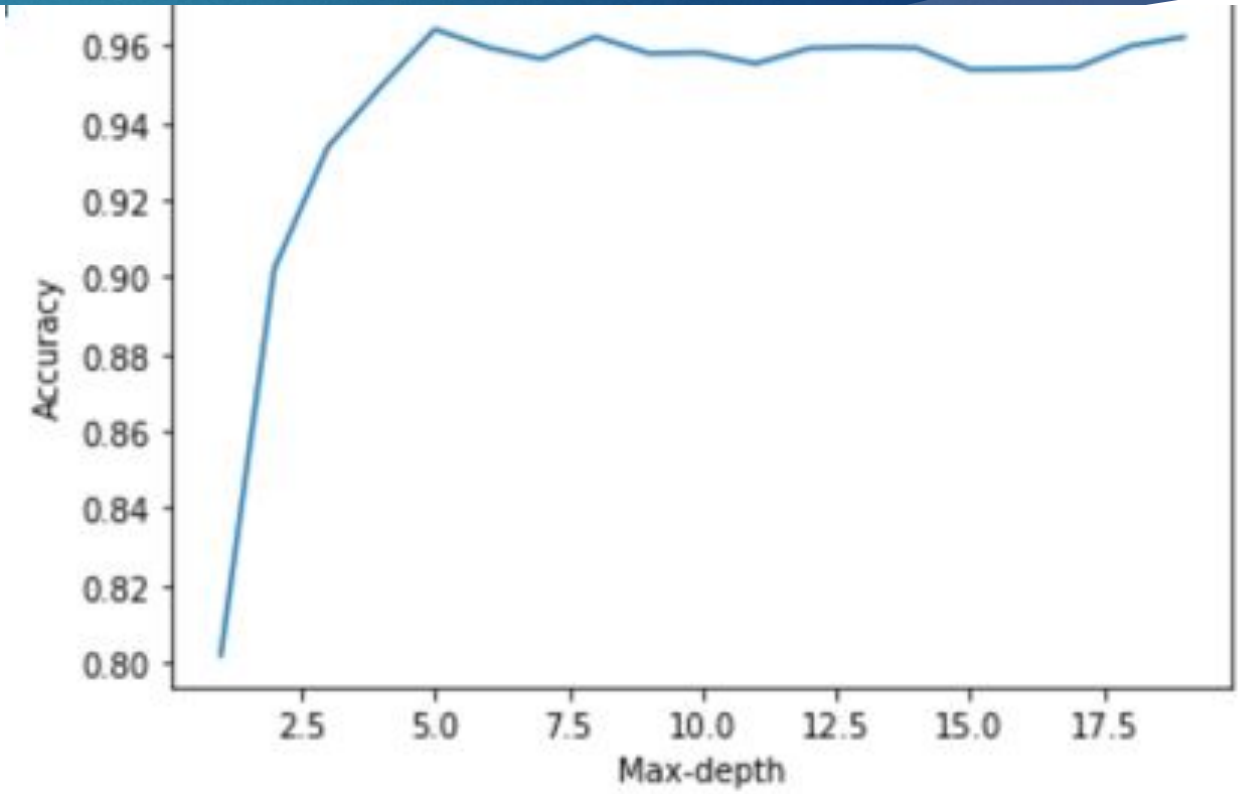


Max depth vs accuracy for cooling load
Max depth = 7 accuracy = 0.959

Fitting random forests on imputed data



Max depth vs accuracy for heating load



Max depth vs accuracy for cooling load

Conclusions from models on imputed dataset.

prediction model	output	r2_score	RMSE	cross validation score
KNN regression	heating load	0.946	2.4369	0.897 at k=5
	cooling load	0.952	2.115	0.910 at k=5
Decision trees Regressor (max depth 7)	heating load	0.996	0.623	0.969
	cooling load	0.944	2.287	0.919
random forest regressor (max depth 5)	heating load	0.996	0.656	0.978
	cooling load	0.963	1.994	0.943

Conclusions from project

Feature	Imputed using features	Selected technique	Accuracy compared to original data
X1	X2 (using mean in X2), Y1	Random forest regressor	0.999
X2	X1, Y1	Random Forest regressor	0.998
X5	X1, X2, Y1, Y2, X4(mode was filled)	•Label Encode •KNN Classifier	0.999
X4	X1, X2, X5, Y1, Y2	KNN Regressor	0.999
X7	X1, X2, X4, X5, Y1, Y2	•Label Encoder •KNN Classifier	1
X3	X1, X2, X4, X5,X6, Y1, Y2	Decision Tree Regressor	1
X6	Y1 and Y2	Decision Tree classifier	0.925
X8	Y1 and Y2	Decision Tree classifier	0.928

Conclusions (Cont…)

- Visualization done using dimensional reduction technique, PCA and found which could be better for the imputed data. Linear regression can not be used as the data was clustered in 12 groups.

Model name	Accuracy on heating load	Accuracy on cooling load
KNN regressor (with optimum K)	0.9043	0.9250
Decision Tree regressor (with optimum depth of the tree)	0.8708	0.9056
Random forest (with optimum depth of the tree)	0.8842	0.9191

Conclusions (Cont…)

- The above three models were fitted on the imputed data along with 10-fold cross validation and finding RMSE.

Model	RMSE	Accuracy	For which output
Decision tree regressor (with depth 7)	0.623	0.9964	Heating Load (Y1)
Random Forest regressor (with depth 5)	1.994	0.963	Cooling Load (Y2)

Learnings from the project:

This analysis can be used to plan and optimize the energy consumption of a new building during its design phase itself.

The kind of model we use for fitting on an appropriate data is largely dependent on the size of data, its distribution and the predictors involved.

The scope of use of machine learning models is vast and it is surprising to find correlations between some of the predictors which do not seem to correlate intuitively.

Importance of machine learning techniques in developing business tactics.

Machine learning can be used to predict the outcomes without performing the task or experiment