Energy Efficiency Prediction using Multioutput Regression Models

BVSS
MSDS
University of Colorado
Boulder

ABSTRACT

Energy efficiency in residential buildings is a multifaceted approach that addresses environmental concerns, economic considerations, and social aspects. Achieving energy efficiency and climate goals requires improvement in energy performance in existing and new buildings. It also plays a crucial role in creating sustainable and cost-effective structures that benefit individuals, communities, and the planet.

This project will give the basic idea of studying heating and cooling loads of distinct types of residential buildings with various parameters. The load calculations will have an impact on energy efficiency.

Multi-output regression involves predicting two or more numerical outputs with a given number of inputs instead of a single-output prediction model avoiding repetitive and time-consuming tasks.

Keywords: Energy-efficiency, Heating and Cooling load, Multi-output regression models.

INTRODUCTION

With increasing population and growing energy demand, efficient use of energy is one of the most cost-effective ways to save money, lower greenhouse gas emissions and helps in improving the wellbeing of the building occupants or homeowners. It also helps the government by providing long-term benefits in lowering the overall electricity demand, thus reducing the need to invest in new electricity generation and transmission infrastructure. By reducing energy consumption and environmental impact, energy efficient buildings contribute to a more sustainable future.

Heating load is the amount of heat energy needed to keep the interior of a building warm during colder periods. In general, the factors influencing the heating load include outdoor temperature, building insulation, windows, doors, and thermal properties of building materials.

Cooling load is the amount of energy required to cool a building and maintain a comfortable temperature during warmer periods. Factors affecting cooling load include outdoor temperature, solar radiation, humidity levels, and the buildings thermal characteristics. Air conditioning units are designed to remove heat from the indoor setting to meet the cooling load requirement.

Multi-output regression models will be used in this study to predict the heating and cooling loads simultaneously for the same set of input parameters.

Understanding and optimizing heating and cooling loads contributes to energy efficiency of a building which is vital in reducing energy consumption and environmental impacts.

The dataset obtained from UCI Machine learning repository consists of 768 samples of twelve different building shapes with eight input features and two real

value responses: heating load and cooling load measured in kWh (kilowatt-hour) regulated by heating ventilation and air conditioning (HVAC) system. The eight input features are:

1. Relative compactness indicator used to show twelve different building types.
2. Surface area measured in square meters (m^2).
3. Wall area measured in square meters (m^2)
4. Roof area measured in square meters (m^2)
5. Overall height measured in meters (m).
6. Glazing area -- 0%, 10%, 25%, 40% of the floor area [0, 0.10, 0.25, 0.40].
7. Orientation – 2 : North, 3 : East, 4 : South and, 5 : West.
8. Glazing area distribution (variance) – [0,1,2,3,4,5] where each number indicates:
   0 : None
   1 : Uniform (25% glazing for each direction).
   2 : North (55% for north face and 15% for each of the remaining three faces).
   3 : East (55% for east face and 15% for each of the remaining three faces).
   4 : South (55% for south face and 15% for each of the remaining three faces).
   5 : West (55% for west face and 15% for each of the remaining three faces).

The first six input features are numerical data types. The remaining two features are categorically encoded data which helps the model to interpret better.

To proceed further it is important to collect relevant data, preprocess it, select appropriate features, choose the regression algorithms (linear regression, knn, decision tree, random forest , multi-output regressor) to train the model and evaluate its performance by evaluating the model with unseen test data. Evaluation metrics like R-squared (R2) , Mean squared error (MSE), Root mean squared error (RMSE), Mean absolute error (MAE) are used to compare the performance of the models and analyze the results.

The five steps flow illustrating Data mining phases is shown in Figure 1.

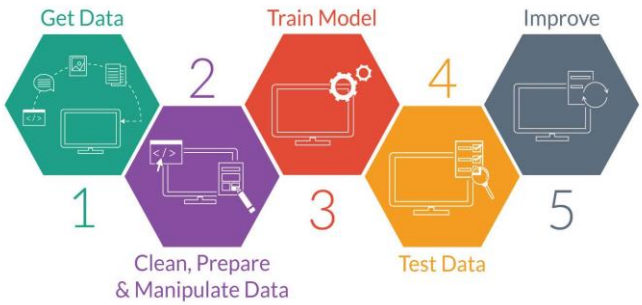


Figure 1 : Data Mining Steps

RELATED WORK

Traditional approaches such as single output regression models have been adapted to predict heating and cooling load separately. Models such as linear regression, knn, decision tree and random forest were used in model building. Here the model is trained twice for each output prediction using the same set on inputs, which is a very tedious and time-consuming repetitive work.

PROPOSED WORK

In this paper, we will build the model using regression algorithms that support multi-output data like linear regression, knn, decision tree and random forest. Support vector regressor (SVR) do not support multi-output data, hence we will propose a multi-output regressor model which is a wrapper model available in sklearn library to use SVR for multi-output data. Data mining and machine learning techniques will be used to find the optimal model the predicts heating and cooling loads in an efficient way. The project will be developed using an open-source web application known as Jupyter notebook which is an interactive computational programming environment.

EXPLORATORY DATA ANALYSIS

Designing an appropriate energy-efficient building depends on its layout such as the eight input features

mentioned in the dataset. All these factors influence the heating and cooling load of the buildings. The dataset taken from UCIML repository consists of 768 instances of twelve different building shapes simulated in Ecotect with various settings as a function of the eight input variables. Ecotect is a building performance simulation software mostly used by architects and building engineers to improve their building designs by calculating the building's energy consumption. Figure 2 shows the graphical representation of how dataset is generated.

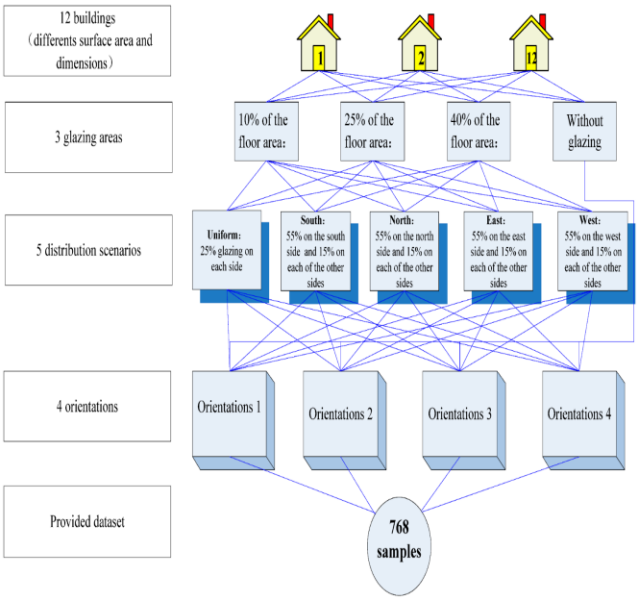


Figure 2 : Graphical view of data generation

After importing the dataset, the most important task in exploratory data analysis is data preprocessing which is done to check the data for missing values and outliers, visualize data distribution to understand and identify any underlying patterns and do correlations to check for relationships among features. There are no missing values or Nan values in the data.

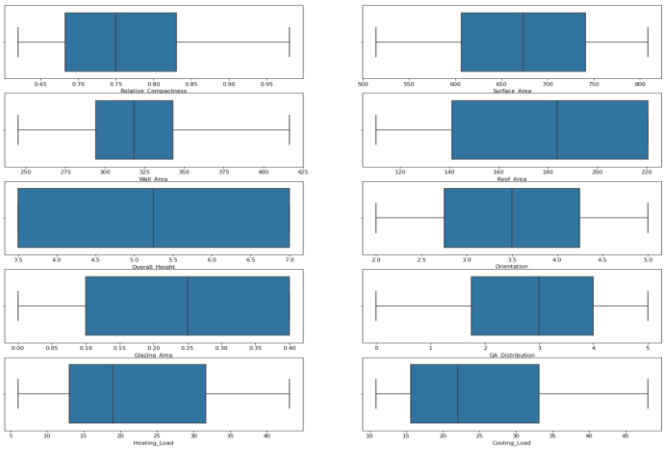


Figure 3 : Box Plot

The boxplot of the data shown above in figure 3 shows there are no outliers present. The input features/variables of the dataset do not follow the gaussian distribution as shown below the figure 4.

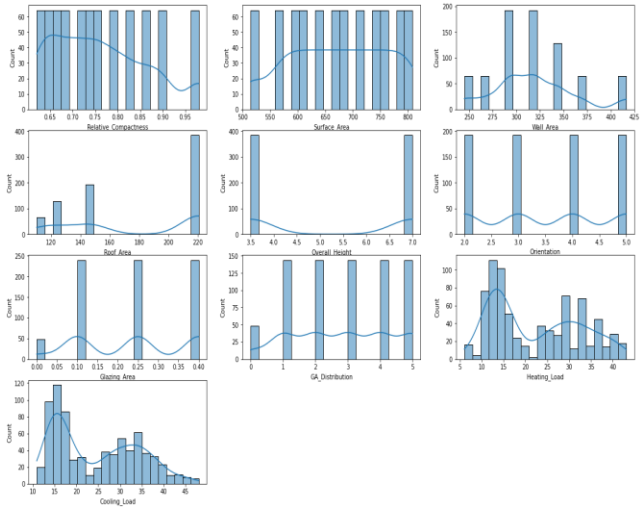


Figure 4 : Data Distribution

Figure 5 below shows correlation between all the variables pairwise.

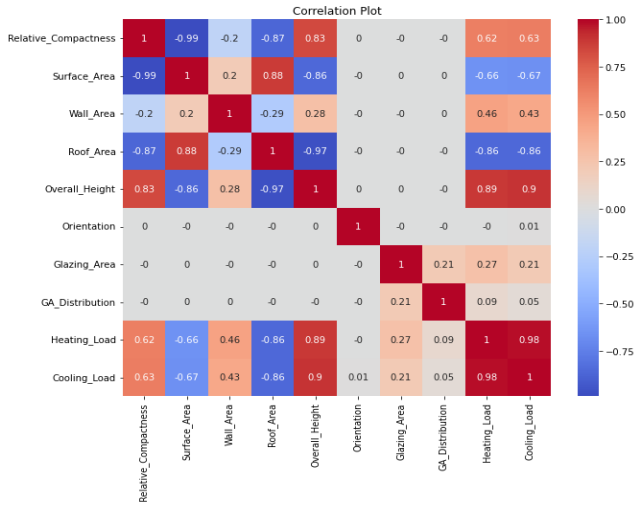


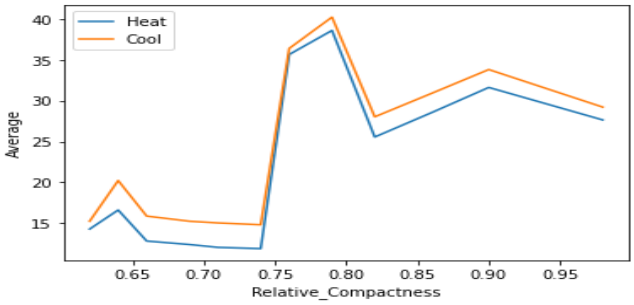
Figure 5 : Correlation Matrix

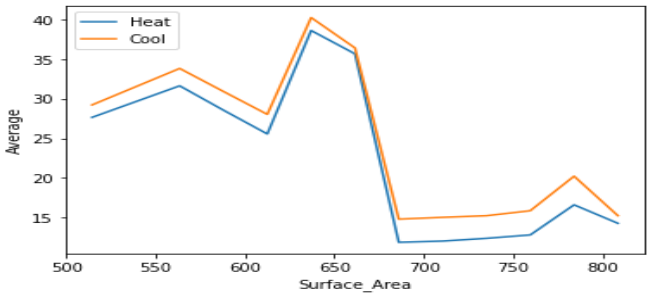
Top Correlated pairs:

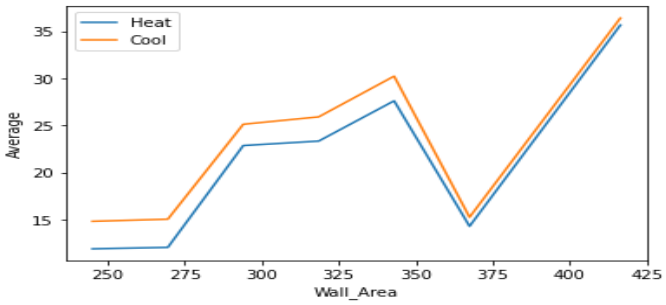
Heating_Load	Cooling_Load	0.98
Surface_Area	Roof_Area	0.88
Relative_Compactness	Overall_Height	0.83

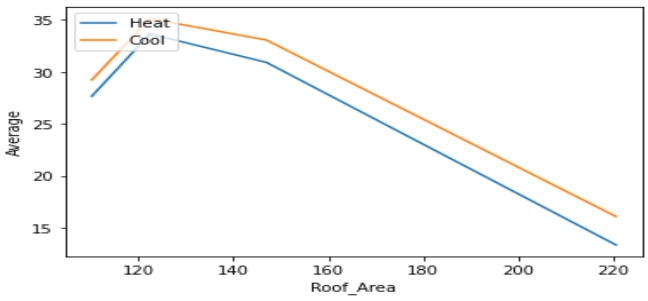
Correlation matrix tells us how strongly the independent variables in a dataset are related to each other. Multicollinearity occurs when two or more independent variables have a high correlation with one another in a regression model. The above data shows existence of multicollinearity between surface area and roof area and between relative compactness and overall height , but since the number of features are less than 10, removing them may cause information loss and an underfitting model. Strong correlation between the target variables heating load and cooling load ensures better predictive performance by using multi-output regressor.

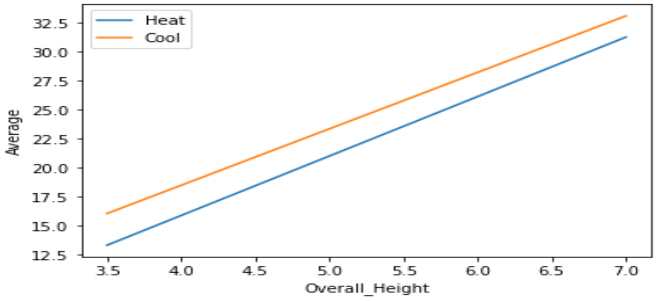
The distribution of average heating and cooling loads with respect to each independent input feature is shown in the below plots. For the same set of input values, the cooling load is higher than heating load and there is strong positive correlation between the two loads.











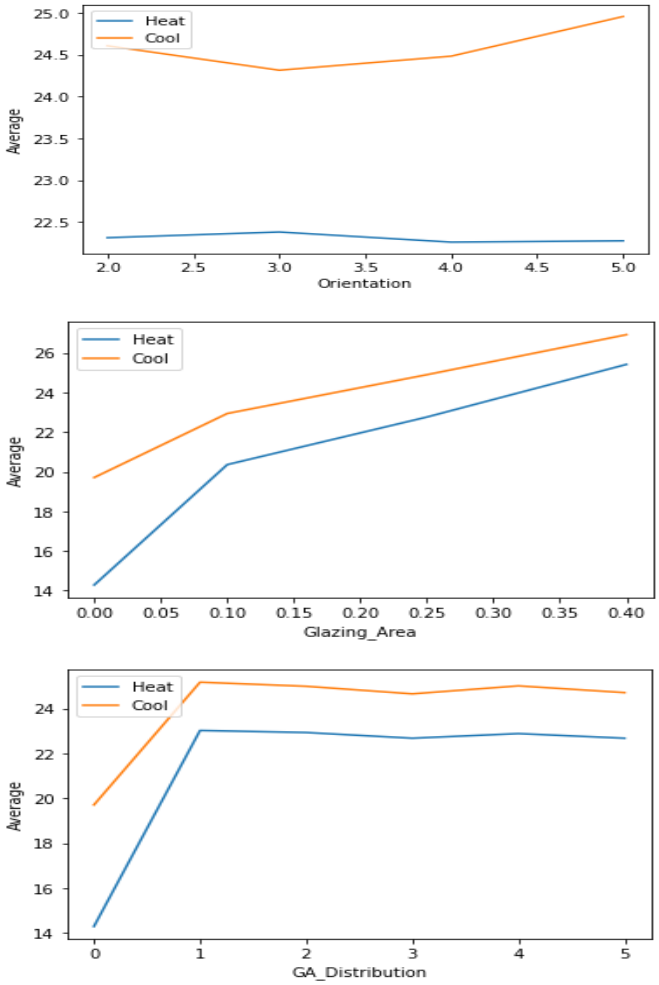


Figure 6 : Heating and Cooling load distributions

DATA TRANFORMATION

When the value of one feature/variable is small as compared to other features in the dataset, then that feature will have not much influence on mining information because of small variation within the feature. Thus, data transformation is a process in which the data is transformed into some other standard forms which are better suited for data mining.

The first five input features in the dataset have different range values i.e., different units and magnitudes of the features may affect the performance of the models due to bias towards numerically larger values. The remaining three columns are categorical values. Feature scaling is one of the most critical steps during the preprocessing stage before building a model with machine learning algorithms. Scaling can have influence between a weak and a better model.

The most common techniques of feature scaling used are Normalization and Standardization while modeling the data using linear regressor or knn regressor model. Since the distribution of the data is non-gaussian and has no outliers, we normalize the data using min-max scaler from sklearn's library after splitting the dataset into training and testing sets (80/20).

DATA MODELING & EVALUATION

The train data is scaled with a min-max scaler and modelled using regression algorithms that support multi-output data.

The models proposed in this study are :

1. Linear Regressor
2. K-Nearest Neighbor Regressor
3. Decision Tree Regressor
4. Random Forest Regressor
5. Multi-output Support Vector Regressor

The first four models directly support multi-output data. Support vector regressor (SVR) does not perform with multi-output and will throw an error while trying to model. Hence, we will be using a wrapper model called multi-output regressor which will extend single-output models like SVR. This wrapper takes input and distributes it into single output regressors that are embedded in it. Predictions generated by the single output regressors are combined and served as multi output regression.

To build a well-generalized model we will evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it and obtain a better result. The evaluation metrics used for regression tasks in this paper are R-squared (R2), mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE). We will

compare the metrics of the five proposed models and analyze the performance.

The R2 value indicates the proportion of variance in the dependent variable that is explained by the independent variables.

The MSE is a measure of quality of an estimator. It measures the average square of errors i.e., the average squared difference between predicted and actual value.

The RMSE is one of the most used statistical parameters in regression models. It shows how much the predicted value differs from the actual value. The lower the RMSE value, the better the model is.

Finally, MAE is another statistical parameter that is used to measure the absolute difference between actual and predicted values. It shows how far the predicted data is from the actual data. The lower the MAE value, the better the model is.

Table 1 and 2 below are the performance scores of all the five models fitted on the train set.

Model	R2	MSE	RMSE	MAE
Linear Regression	**0.897**	**9.888**	**3.145**	**2.234**
Knn	**0.968**	**3.083**	**1.756**	**1.253**
Decision Tree	**1.000**	**0.000**	**0.000**	**0.000**
Random Forest	**0.997**	**0.300**	**0.548**	**0.282**
Multioutput SVR	**0.994**	**0.562**	**0.750**	**0.376**

Table 1: Model Performance on raw data.

Model	R2	MSE	RMSE	MAE
Linear Regression	**0.897**	**9.895**	**3.146**	**2.235**
Knn	**0.929**	**6.901**	**2.627**	**1.791**
Decision Tree	**1.000**	**0.000**	**0.000**	**0.000**
Random Forest	**0.997**	**0.305**	**0.552**	**0.287**
Multioutput SVR	**0.910**	**8.625**	**2.937**	**1.821**

Table 2: Model Performance on scaled data

Comparing the results of Table 1 and 2 yields interesting information.

Modeling on scaled data has no effect on the performance of Linear Regressor.

Since Decision Trees and Random Forest are tree-based models they are unaffected by scaling.

KNN and Multi-output SVR models had a better score on raw data than scaled data.

Ideally scaling enhances the model's performance but due to the number of instances in the dataset being fewer, it did not have much impact on the model performance. If we have a greater number of instances in the dataset, scaling will increase the model accuracy. Also, the dataset contains both numerical and categorically encoded data which often confuses the model when scaling is performed. However, not all algorithms benefit from feature scaling. Normalization can sometimes amplify noise or irrelevant variations in the data which may negatively affect the model's performance.

The relationships between the target variables and the input features are non-linear which caused scaling to not improve the model's performance. Hence in this study, performance of the regression algorithms is compared using raw data since min-max scaling is not the most appropriate scaling method for this data set.

Table 3 below shows the scores of all the five models fitted on the test data.

Model	R2	MSE	RMSE	MAE
Linear Regression	0.896	9.124	3.021	2.157
Knn	0.921	7.003	2.646	1.925
Decision Tree	0.977	1.933	1.390	0.629
Random Forest	0.981	1.585	1.259	0.659
Multioutput SVR	0.984	1.417	1.190	0.791

Table 3: Model performance on raw test data

Decision tree model outperforms all other models on train set. Random Forest and Multi-output SVR did well too. The MSE, RMSE and MAE scores of these 3 models are closer to 0.

Multi-output SVR has the best test score with 0.984 (R2 in the above table) with better MSE,RMSE and MAE values which indicates a good model. The scores of Decision tree and Random Forest models are also good.

Table 4 below shows the R2 scores of single-output and multi-output models evaluated on the test data.

R2_MO : Multi-output model R2 score
R2_SO_HL: Single-output model R2 score of Heating Load
R2_SO_CL: Single-output model R2 score of Cooling Load

Model	R2_MO	R2_SO_HL	R2_SO_CL
Linear Regression	0.896	0.910	0.890
Knn	0.921	0.904	0.890
Decision Tree	0.977	0.997	0.950
Random Forest	0.981	0.998	0.962
Multioutput SVR	0.984	None	None
SVR	Not supported	0.911	0.887

Table 4: Test data Single-output and Multi-output model comparison table

HYPERPARAMETER TUNING

Hyperparameter tuning is essential for optimizing model performance, improving generalization, and ensuring the models are well-suited for their specific tasks.

The performance of Linear Regression and Knn regression models can be improved with hyperparameter tuning. GridSearchCV is a popular library in Python's Scikit-learn module. It performs an exhaustive search over a specified parameter grid

to find the set of hyperparameters that yields the best performance for a given model using cross validation.

Hyperparameter tuning of Linear regression model did not lead to improved model performance. This may be due to insufficient data size.

Knn regressor model scores improved with hyperparameter tuning. The train and test score of the model slightly improved to 0.975 and 0.934 respectively compared to the base model's train and test score of 0.968 and 0.921.

Hyperparameter tuning is not performed on Decision tree, Random forest, and Multi-output regressor models since they have satisfied scores.

The plot of the actual vs predicted values for the train and test set of all 5 models is shown in the plot below in Figure 7.

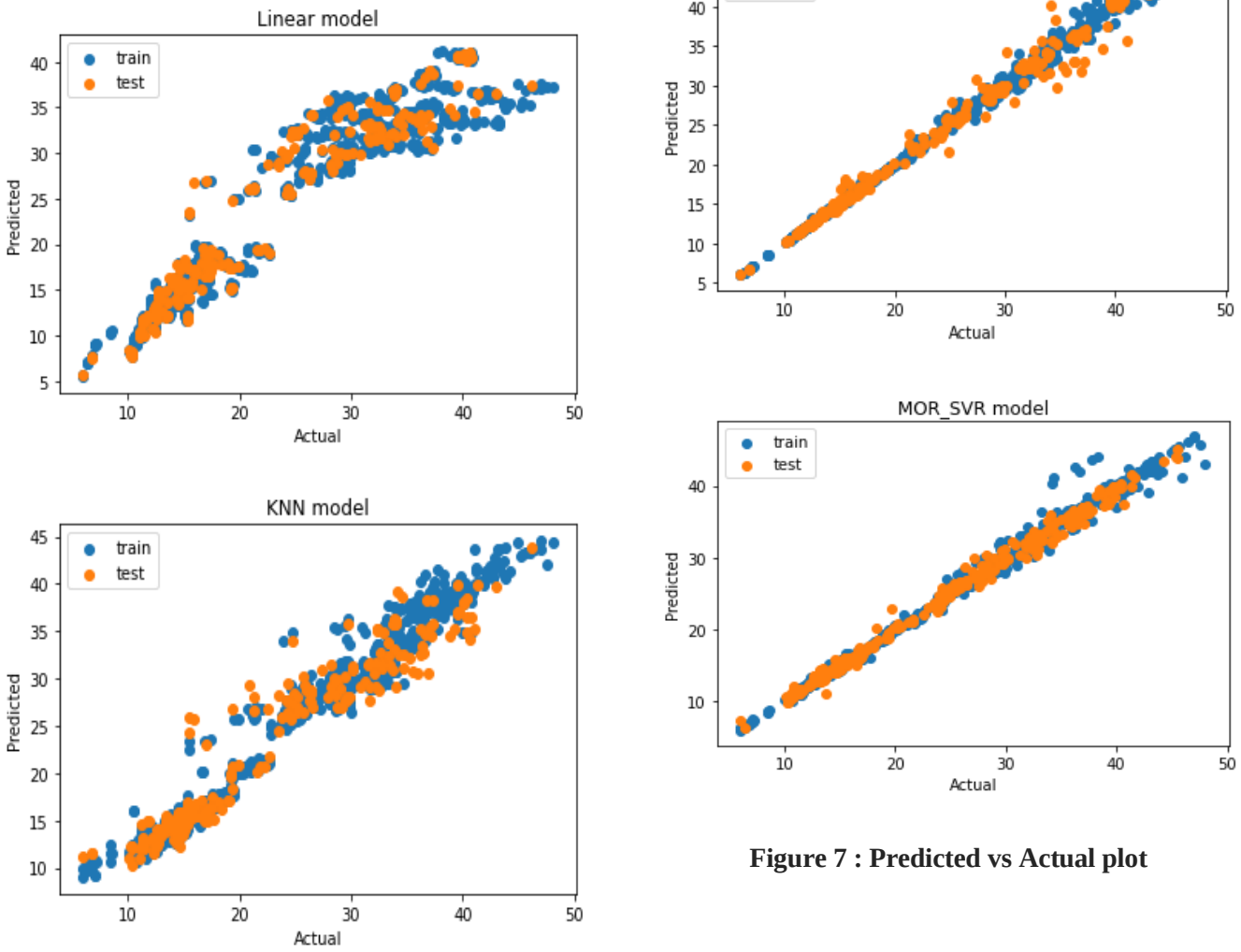


Figure 7 : Predicted vs Actual plot

CONCLUSION

Overall, the results of this study provided valuable insights on the effectiveness of different regression algorithms on predicting heating and cooling loads which can be helpful in improving energy efficiency of buildings and reduce energy costs.

Multi-output models' performance is better than separate single-output models, hence can be used to avoid repetitive work.

Overall, the Multi-output SVR model performed better with R2 score of 0.984, mean square error of 1.417, root mean square error of 1.190 and mean absolute error of 0.791 than other four models.

With hyperparameter tuning the performance of KNN regressor model slightly improved. The plot of actual vs predicted values of both train and test on the best model obtained after tuning is shown below in Figure 8.

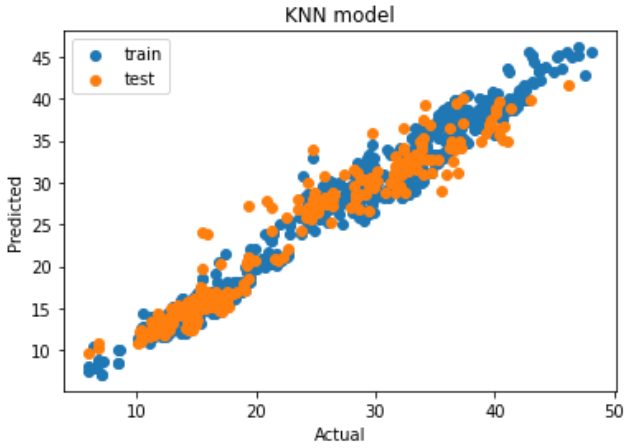


Figure 8: Predicted vs Actual plot (Tuned Knn)

FUTURE WORK

Increasing the size of the data and analyzing with Deep learning models is intended as future work.

TIMELINE

1. Read the data and perform basic exploratory data analysis. (Week 1)
2. Data Preprocessing- Feature Scaling. (Week 1) - Done
3. Split the data into the train and test sets. (Week 1)
4. Model the data using train set. (Week 2) - Done
5. Choose the best model and evaluate on tests set. (Week3)
6. Check the performance with hyperparameter tuning. (Week 4)
7. Compare the results and report final conclusions. (Week 4)

REFERENCES

1. "Robust modeling of heating and cooling load using partial least squares towards efficient residential building design." July 2018, https://www.sciencedirect.com/science/article/abs/pii/S2352710218301128

2. "Towards efficient building designing: heating and cooling load prediction via multi-output model.", published online 10 November 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7696299/

3. "Forecasting heating and cooling loads in residential buildings using ML." 28 July 2023, https://link.springer.com/article/10.1007/s42107-023-00834-8

4. "Predicting heating load in energy efficient buildings through ML techniques." 15 October 2019, https://www.mdpi.com/2076-3417/9/20/4338

5. "Building heating and cooling load prediction using ensemble learning model.", 10 October 2022, https://www.researchgate.net/publication/364302251_Building_Heating_and_Cooling_Load_Prediction_Using_EnsembleMachine_Learning_Model