**Research Paper**

INT247

Submitted By:

Satyam Chand Deopa

11806590

RKM001B53

Submitted To:

Usha Mittal

**Predicting Energy Efficiency of Homes**

This paper contains many models to make a comparison and find the best model with highest accuracy. Prediction of Heating and Cooling Loads is the key.

GitHub Link:

**Abstract**

Machine learning methods can be used to help design energy-efficient buildings reducing energy loads while maintaining the desired internal temperature. They work by estimating a response from a set of inputs such as building geometry, material properties, project costs, local weather conditions, as well as environmental impacts. These methods require a training phase which considers a dataset drawn from selected variables in the problem domain. This paper evaluates the performance of many machine learning methods to predict cooling and heating loads of residential buildings. The proposed framework resulted in accurate prediction models with parameters that can potentially avoid modeling and testing various designs, helping to economize in the initial phase of the project.

**Introduction:**

**Intro to the problem**

This paper focuses on the efficiency depicted by housing building. The energy required to warm/cool is affected by the appliance used and the materials it consists of. However, we look at a different aspect of building. Its dimensions, the roof, floor size and the perimeter(walls). To use this functionality, we are using the Building Energy-Efficiency Dataset, available on UCI repository under the name Energy-Efficiency.

**About Dataset**

The dataset consists of 768 samples with eight input attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses. It contains data about residential houses: Their planning and Surface Areas and how much is of it is covered in shade or has direct sunlight. The main objective is to predict the heating and cooling loads required in these homes.

Specifically:

X1 - Relative Compactness

X2 - Surface Area

X3 - Wall Area

X4 - Roof Area

X5 - Overall Height

X6 - Orientation

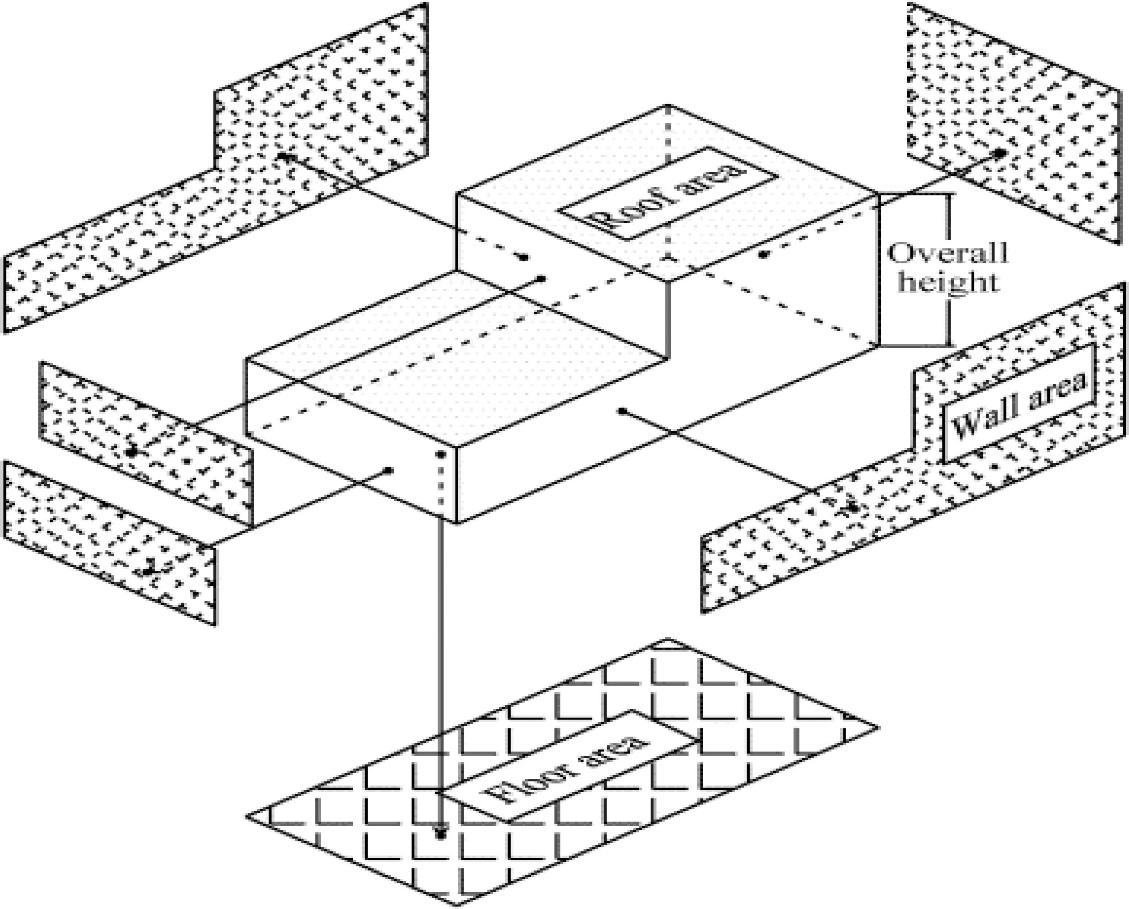
X7 - Glazing Area

X8 - Glazing Area Distribution

y1 - Heating Load

y2 - Cooling Load

Some inspection on the data reveals that there are no missing values in the dataset and hence do not need to impute any value within the dataset.



**Limitations of the already existing technology**

Many of the already existing technologies use different kinds of functions. Some of the technologies are using quite complex functionalities but still not resulting equivalently high accuracies.

**My Contribution**

The goal of this paper is to use the dataset and create a model which can predict the heating/cooling load with high accuracy while using a relatively simpler model. We used many models that support regression, including SVM, Linear and Logistic Regression, Lasso and Polynomial Regression. This was achieved using polynomial regression, which gave the best results among the different models that we used.

**Literature Survey**

The first paper focuses primarily on bagging and boosting algorithms, along with a few more complex algorithms and some linear models. However, the best results were achieved by the Random Forest Regressor, with a 71 & 59% accuracy on Y1 & Y2 respectively.

The second paper discussed the various types of models and possibilities that can be utilized on the dataset. The model of focus was a custom-made ANN with 3 hidden layers. The dataset was however not split for validation data and was trained for 3000+ cycles. The model resulted an accuracy of 99%. However, chances of overfitting of the model is quite high for the same reason and the model may not give a good result when exposed to unseen data.

The third paper focuses on a basic Naïve Bayes Classifier. It gets an accuracy of 83%.

The fourth paper has great data visualization and also gives and explains the methods and models used to predict the Heating and Cooling Load. It consists of many models ranging from basic to more complex ones. Apart from the linear, and lasso regression, the author also uses various gradient boosting methods such as GBM, XGBoost, and LightGBM. Wherein all the gradient boosting methods give a r2 score of 99% for both loads.

The fifth paper had an accuracy of 99% and 94% on predicting Heating Load and Cooling Load, many models were considered RF[3], EMARS [7], SVR with ANN[5], GP[4], MLPwith grid search strategy [6]and RBFNN. Best working were RBFNN and GP.

Paper sixth takes a similar approach and train models with K-fold method. Where the validation set keeps rotating every iteration. The same question arises whether the model has made conclusive inferences and will be able to predict loads near the actual target.

The seventh paper took quite a different approach and delved into fuzzy logic to predict the loads. The model gave an RMSE of 1.2 and 1.75 on heating and cooling loads respectively.

The eighth paper also gave a great insight over the data visualization and correlations between the attributes. The models used were however genetic/evolutionary and hence required much more time before reaching a conclusion. PSO, GA, ICA, ABC, GA-ANN. The best r2 score was 98% for GA-ANN.

The ninth paper was focused on quick algorithms as well like ANN, SVM, RVC. However, the highest score achieved was with ANN of 86%.

The tenth paper was based on deep learning models and the best results were achieved with a DELM(Deep Extreme Learning Model). Other models considered were ANFIS and ANN, with ANFIS predicting better than ANN.

The eleventh paper used different kinds of models to test on the dataset. However, they failed to get good results and predict accurately.

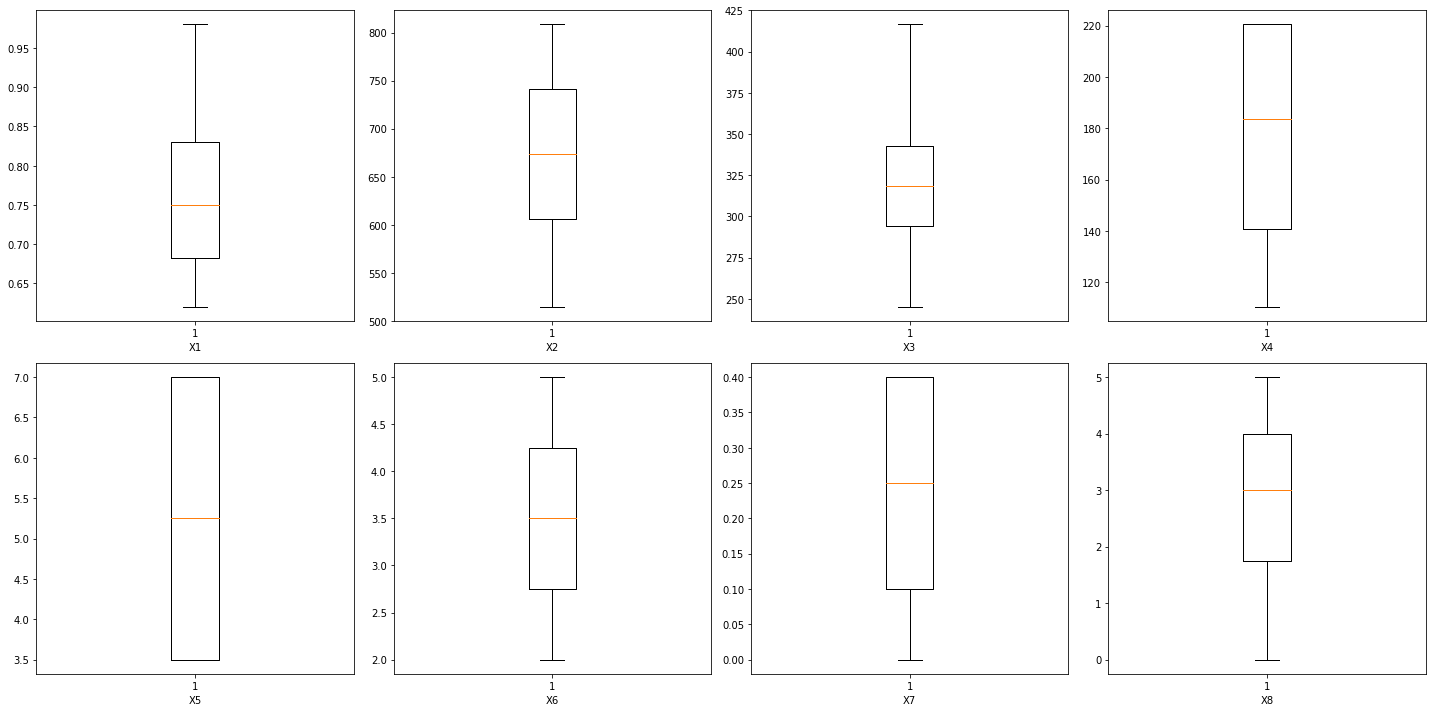
**Proposed Methodology**

Through the duration of this paper, we will try to achieve high accuracies while trying to keep the models relatively simplistic (hence will take less time to train, test) with a bit of Data transformation. Most of the papers reviewed during this paper had heavy models, or the models which are quite complex and are not known to beginners. The others which utilized a simplistic approach like Linear/Logistic Regressors, SVM did not handle the data very well and failed to give higher accuracies.

We are going to explore many regressors including Liner Regressor, Logistic regressor, Random Forest Regressor etc. Also, we will try to find the best model utilizing data transformation and see the difference in polynomial regression after increasing the dimensionality. First, we will try to train models on regular dataset without any tempering and further see if adding or removing dimensionality, columns help us in finding the best model. For this we can use Data visualization to factorize which columns to keep, find how the columns are set in respect to each other using correlation and just know more about our data.

**Data Visualization**

First, we try to visualize the dataset and see some basic factors in our data like dispersion of data and correlation of columns with each other and their impact on the output fields. The dispersion of data can be visualized by plotting boxplots and get a sense of scale for different columns. These are displayed down below:



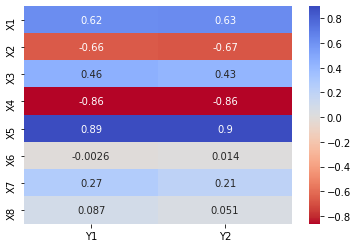
The general inclination when finding outliers in data is to impute them and make data well within a range. However, when imputed for the given dataset, it reduces the accuracy which means that the data in the outlier range is also significant in our case and helps us reach a better conclusion and hence we cannot impute them.

Next, we will like to how our columns correlate to each other and how they stand in relation to each other. We can easily visualize this by finding the correlation of the data and plotting it into a heatmap. This is shown as below.



The brighter the blue color represents that the column is more similar to each other. Similarly, Brighter red means they have a reverse affect i.e., if value of one column increases, the other decreases.

However, more than how the columns stand with each other, we are interested in how the different input columns affect our output. This can be visualized as below:



As we can infer from the above heatmap, the different inputs have different effects on our output and even though we have some correlations of similar intensity, they are on the opposite end of the scale. And from the graph we can safely conclude that dimensionality reduction will have any appreciable affect towards increasing our accuracy.

**Models**

**Linear**

A simple linear model helps us set an idea of what we should expect for the dataset that we are using. Although it is quite known that no model is perfect for every data, setting up how other linear models might do can help us get an estimate how we are going to do. The linear model we created was trained directly on the dataset without any normalization/regularization. The model yielded a test accuracy of 0.8993. A model was also tested on a normalized dataset; however, it returned a lower accuracy.

**Polynomial**

The polynomial model works on adding dimensionality to the data and allows a Linear model to also predict for a curved estimator. For example, of we need to have a curved regression line on the dataset, we can increase the dimensions and further train a linear model to achieve desired results.

For this dataset, we utilized Linear Regression on top of modified data and models waere tested on many degrees. The most significant ones of them were:

2nd Degree:

The dataset was transformed into 2-degree dataset to allow us test the possibility of a curved regression giving a better output. This came out to be true as the model showed accuracy of 0. 9773, which is much higher than the Simple Linear regression.

3rd Degree:

Since the adding of dimensionality appeared to be quite beneficial, we continued to add another dimension and reached 3rd degree in the dataset. The results were even better than the previous with an accuracy of 0.9829, which is the best so far.

* Further degrees were also taken into consideration but the accuracies achieved were either not better or weren’t significantly higher even after adding a lot of dimensions.

**Ridge**

Ridge regression is a model which performs regularization by itself. It performs L2 regularization, i.e., adds penalty equivalent to square of the magnitude of coefficients. The accuracy achieved by this model was 0.8984 even after tuning the value for alpha parameter.

**Lasso**

Similar to the Ridge model, Lasso regression is also a regularization technique. However, it performs L1 regularization, i.e., adds penalty equivalent to absolute value of the magnitude of coefficients. The accuracy achieved through this model was 0.8952.

**ElasticNet**

ElasticNet is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. The L1 and L2 regularizations can be set in a preferred ratio to balance the regularizations and make the most out of this model. The accuracy achieved through this model was 0.8242. However, no parametric tuning was done in this case.

**Random Forest**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. Accuracy achieved using Random Forest Regressor was 0.9775.

**Decision Tree**

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. The model worked really well in the first run and hence we went on a bit ahead with it to check possibilities.

Since the accuracy of Decision tree was comparable to the best found so far (i.e. Polynomial 3D), We tuned it and found that it was giving the best result with min\_impurity\_decrease = 0.043 to 0.045. However, the accuracy was fluctuating much and was still less than the accuracy for Polynomial 3D.

The accuracies in the final take were 0.9649 and 0.9795 for untuned and tuned Decision Tree respectively.

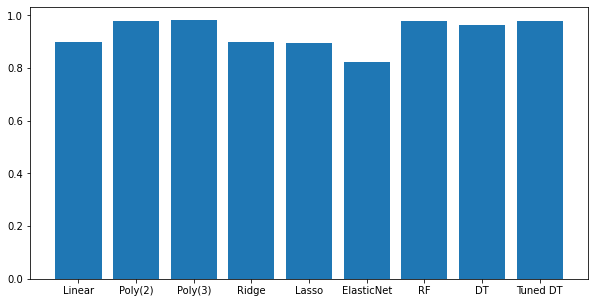
**Results and Discussion**

The accuracies of different models are shown in the below table.

|  |  |
| --- | --- |
| Model | Accuracy |
| Linear | 0.8992 |
| Polynomial (2nd Degree) | 0.9773 |
| Polynomial (3rd degree) | 0.9829 |
| Ridge | 0.8984 |
| Lasso | 0.8952 |
| ElasticNet | 0.8242 |
| Random Forest | 0.9775 |
| Decision Tree | 0.9649 |
| Tuned Decision Tree | 0.9795 |

The highest accuracy was noticed for 3rd degree polynomial. It was possible with tuned decision tree to almost match the highest accuracy but it varied a lot. Random Forest and 2nd Degree Polynomial were also close to the top accuracy achieved.

A bar graph with visual representation is also down below.



**Conclusion and Future Scope**

We can conclude from our findings that adding dimensionality and making the dataset 3rd Degree Polynomial yielded the highest accuracy of 0.9829 or 98.29% overall.

Machine learning techniques have become very popular in recent years. Machine learning techniques have become a very popular tool in recent years due to their potential capabilities, as they can be used to solve for their flexibility and adaptability, as well as scalability for different practical applications, i.e., complexity of machine learning tools in use, can be modulated depending on the requirements of the problem to be solved.

The paper presented aims to solve and validate the problem of load generations for heating and cooling residential buildings based on their dimensions and some other factors. We have introduced the results of the research carried out focusing on the use of machine learning, in the field of the energy consumption of buildings. Among other possibilities, we highlight the potential use in layout planning of future homes that will require less loads to keep the home at an appreciable temperature during both summer and winter. This can help residents further in saving electricity and saving efficiently.

The findings and model can be further improved by increasing the dataset or adding more functionality as it is known that the loads required to maintain the temperature is not only based on the layout but also the materials used on different surfaces. However, this will be a point of improvement of dataset. The model can also be possibly made better if along with the addition of dimensions (as shown in the paper), we use some more sophisticated models, preferably of non-linear models to achieve even better results.

**References**

urnal is © Advanced Technology & Science IJISAE, 2017, 5(2), 81–85 | 81

1. Comparison of Classification Techniques on Energy Efficiency Dataset Ahmet TOPRAK, Nigmet KOKLU, Aysegul TOPRAK, Recai OZCAN

<https://www.researchgate.net/publication/322001360_Comparison_of_Classification_Techniques_on_Energy_Efficiency_Dataset>

1. Load Finding Using K-fold ANN

<https://core.ac.uk/download/pdf/237182408.pdf>

1. Analysis of building energy efficiency dataset using naive bayes classification classifier to cite this article: B Prasetiyo et al 2019 J. Phys.: Conf. Ser. 1321 032016

<https://iopscience.iop.org/article/10.1088/1742-6596/1321/3/032016/pdf>

1. Comparison of machine learning techniques for predicting energy loads in buildings

Grasiele Regina Duarte[1](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103#aff1)Leonardo Goliatt da Fonseca[1](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103#aff1)Priscila Vanessa Zabala Capriles Goliatt[1](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103#aff1)Afonso Celso de Castro Lemonge[1](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103#aff1)

<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103>

1. Prediction and Analysis of Building Energy Efficient Neural Network and Design of Experiments Sholahudin, Azimil Gani Alam, Chang-In Baek

<https://www.researchgate.net/publication/289554036_Prediction_and_Analysis_of_Building_Energy_Efficiency_Using_Artificial_Neural_Network_and_Design_of_Experiments>

1. Prediction of Heating and Cooling Load to improve Energy Efficiency of Buildings Using Machine Learning Techniques 1 Srihari J, 2 Santhi B

<http://www.journalimcms.org/wp-content/uploads/8-Prediction-of-Heating-and-Cooling.pdf>

1. Estimation of Building Energy Efficiency Performance Using Radial Basis Function Neural Network Che Munira Che Razali\*,Shamsul Faisal Mohd Hussein2, Nolia Harudin3, Shahrum Shah Abdullah4

<https://www.sciencepubco.com/index.php/ijet/article/view/23102/12059>

1. Prediction of energy load of buildings using machine learning methods Grasiele Regina Duartea1 , Priscila Caprilesa Afonso Lemongea and Leonardo Goliatta

<http://www.epacis.net/ccis2016/papers/paper_69.pdf>

1. Fuzzy Model for Estimation of Energy Performance of Residential Buildings Shirley Urdaneta1 ,Emilio Zapata1 andJuan Contreras1,2, a

<https://www.ripublication.com/ijaer17/ijaerv12n11_20.pdf>

1. A Comparative Study of PSO-ANN, GA-ANN, ICA-ANN, and ABC-ANN in Estimating the Heating Load of Buildings’ Energy Efficiency for Smart City Planning by Le Thi Le1,\*, Hoang Nguyen2,\*, Jie Dou3 and Jian Zhou4

<https://www.mdpi.com/2076-3417/9/13/2630>

1. Comparison of Classification Techniques on Energy Efficiency Dataset

A. Toprak, Nigmet Koklu, +1 author Recai Ozcan

Published 2017

<https://www.semanticscholar.org/paper/Comparison-of-Classification-Techniques-on-Energy-Toprak-Koklu/d6f008340d47715ac65c837b9ae56dcebd1a7cd3>

And a few other papers were used for refence as well but they were based upon different but slightly related topics.