# Energy Efficiency Prediction

## Introduction

This research evaluates building heating and cooling load requirements, focussing on energy efficiency as a function of several building parameters. Our target variable, the average energy load of a structure, will be the focus of this study to understand its components. We use a variety of statistical modelling methods to gain insights that help improve building energy efficiency.

## Data Preparation

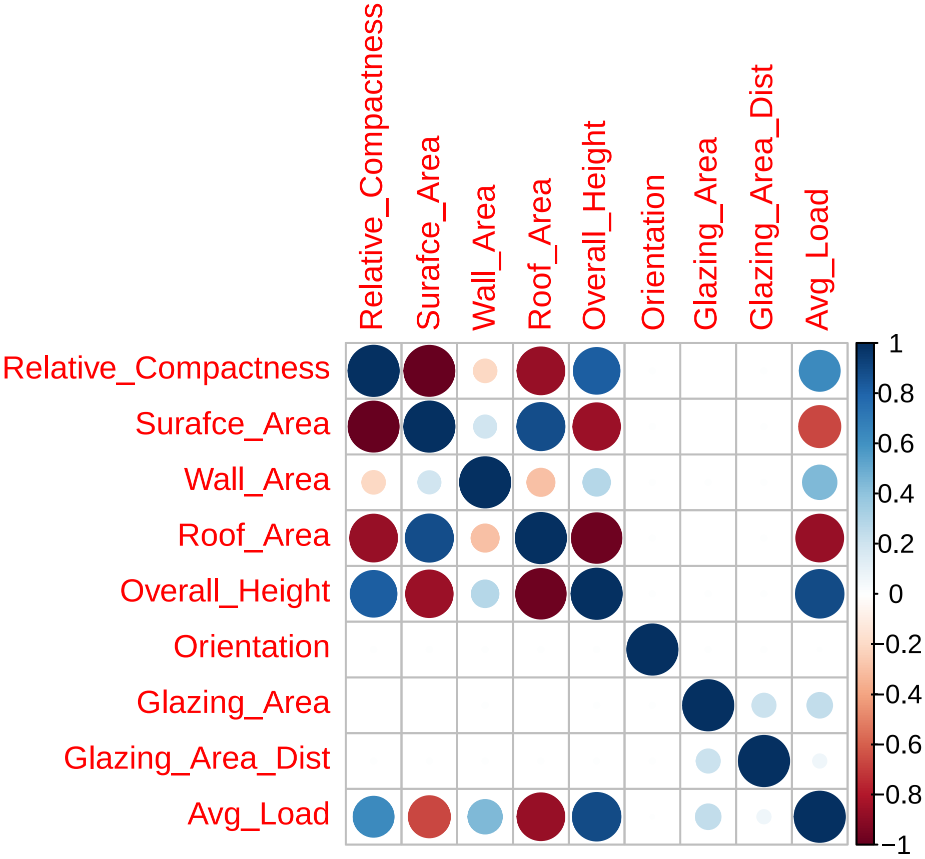
The methodology of this study is comprised of a number of important elements, the first of which is a comprehensive understanding and preparation of the data. There are several other characteristics that pertain to the design of buildings that are included in the collection. These characteristics include Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, and Glazing Area Distance. In the course of the phase devoted to the preparation of the data, it was verified that the dataset did not contain any missing values, outliers, or data type incompatibilities. An additional target variable was developed as part of the analysis. This new variable represents the average of the heating load and the cooling load. In order to simplify the study, the variables that were originally used to represent the heating load and the cooling load were eliminated.

```
## Relative_Compactness         Surafce_Area            Wall_Area
##                    0                    0                    0
##            Roof_Area       Overall_Height          Orientation
##                    0                    0                    0
##         Glazing_Area    Glazing_Area_Dist         Heating_Load
##                    0                    0                    0
##         Cooling_Load
##                    0
```

According to the summary statistics of the dataset, the average load varies greatly, with the maximum recorded average load reaching 44.975. This indicates that the average load is highly variable.

```
##  Relative_Compactness  Surafce_Area     Wall_Area       Roof_Area
##  Min.   :0.6200       Min.   :514.5   Min.   :245.0   Min.   :110.2
##  1st Qu.:0.6825       1st Qu.:606.4   1st Qu.:294.0   1st Qu.:140.9
##  Median :0.7500       Median :673.8   Median :318.5   Median :183.8
##  Mean   :0.7642       Mean   :671.7   Mean   :318.5   Mean   :176.6
##  3rd Qu.:0.8300       3rd Qu.:741.1   3rd Qu.:343.0   3rd Qu.:220.5
##  Max.   :0.9800       Max.   :808.5   Max.   :416.5   Max.   :220.5
##  Overall_Height  Orientation    Glazing_Area    Glazing_Area_Dist
##  Min.   :3.50   Min.   :2.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.:3.50   1st Qu.:2.75   1st Qu.:0.1000   1st Qu.:1.750
##  Median :5.25   Median :3.50   Median :0.2500   Median :3.000
##  Mean   :5.25   Mean   :3.50   Mean   :0.2344   Mean   :2.812
##  3rd Qu.:7.00   3rd Qu.:4.25   3rd Qu.:0.4000   3rd Qu.:4.000
##  Max.   :7.00   Max.   :5.00   Max.   :0.4000   Max.   :5.000
##     Avg_Load
##  Min.   : 8.475
##  1st Qu.:14.375
##   Median :20.485
##   Mean   :23.447
##   3rd Qu.:32.167
##   Max.   :44.975
```

The overall height revealed a high positive association with the average load, whereas the surface area and roof area both exhibited large negative relationships with the load. This latter finding is particularly noteworthy. The Wall Area variable displayed a weak positive association with the average load, however the other variables did not reveal any meaningful correlations with one another.



## Machine Learning Models

R implemented Linear Regression, Decision Trees, and Random Forests for analysis. Linear Regression was chosen for its simplicity and interpretability to understand predictor elements and average load. Decision Tree data modelling captures complicated variable interactions and is more flexible and non-linear. Multiple decision trees were averaged in Random Forests to improve prediction accuracy and avoid overfitting. Cross-validation tuned Decision Tree and Random Forest hyperparameters to improve model performance. Each model was assessed using MAE, MSE, and R Square.

## Performance of Models

After the research, several surprising patterns and significant performance improvements across modelling approaches were found. The Linear Regression model has a Mean Absolute Error of 1.88, a Mean Squared Error of 6.70, and a R Square value of 0.92, indicating a good fit. These values imply acceptable model accuracy. The Decision Tree model performed much better. The mean absolute error was 1.40, the mean squared error was 3.34, and the R Square value was 0.96, indicating that it accurately captured the data's relationships. The Random Forest model improved predicted accuracy with an MAE of 0.49, MSE of 0.61, and R Square of 0.99. These results show that the model can make virtually flawless predictions and that this ensemble technique works.

```
## note: only 7 unique complexity parameters in default grid. Truncating the grid to 7 .
```

```
## [1] "Model Performance Summary:"
```

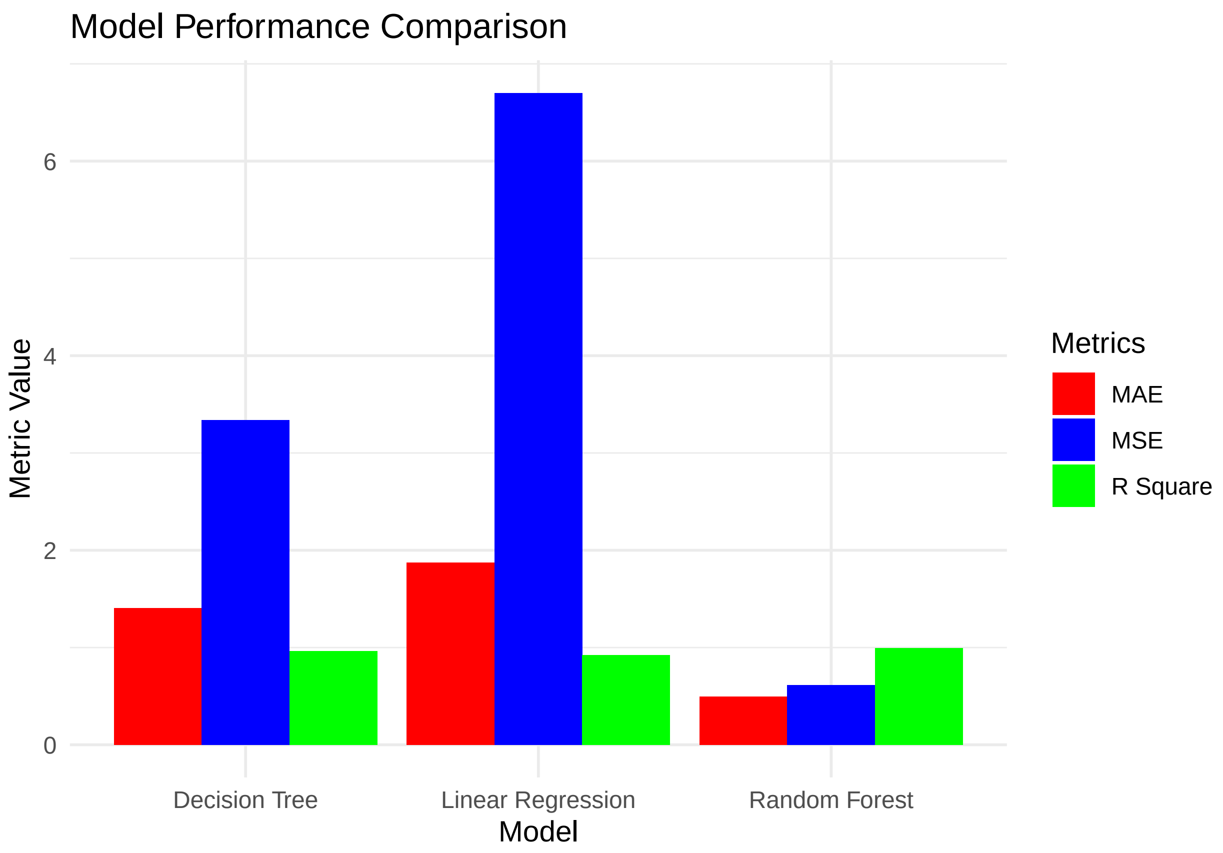
```
##               Model      MAE       MSE R_squared
## 1 Linear Regression 1.875271 6.7020655 0.9249281
## 2     Decision Tree 1.403253 3.3391230 0.9623486
## 3     Random Forest 0.494520 0.6141339 0.9932106
```

## Comparison of Models

The Linear Regression model is a baseline because of its simplicity and interpretability. It posits a linear relationship between predictor factors and target variables, revealing how each component affects average load. The Linear Regression model had an MAE of 1.88, an MSE of 6.70, and a R Square of 0.92. These results show a good match to the data, but the model's linear assumptions may not capture the intricacies of variable connections, especially interactions.

Decision Tree data modelling is more flexible and non-linear. The Decision Tree may detect and express complicated interactions and nonlinear relationships between variables by segmenting the data by attributes. Decision Tree results indicated a significant improvement in prediction performance, with an MAE of 1.40, MSE of 3.34, and R Square of 0.96. This approach captures data patterns better than Linear Regression since it handles variable interactions. Decision Trees can overfit, especially when improperly calibrated.

The Random Forest model, which includes several Decision Trees, outperforms Linear Regression and Decision Trees. The Random Forest had excellent prediction accuracy with an MAE of 0.49, MSE of 0.61, and R Square of 0.99. This increase is due to its capacity to average numerous tree predictions, lowering Decision Tree variation. Random Forests are less likely to overfit since they randomly select tree features, which improves generalisation to new data. For this higher accuracy, Random Forests are harder to read than Linear Regression since knowing their characteristics is more complicated.



## Conclusion

Finally, this study showed how architectural features affect energy load requirements. Thus, it illuminated building energy efficiency factors. As shown by the Random Forest model's better performance than Linear Regression and Decision Trees, adopting appropriate modelling methodologies is crucial. These insights can assist architects and engineers construct energy-efficient buildings by focussing on key characteristics that affect energy loads. Research may find strategies to improve predictive models. To improve energy load calculations, architectural components and powerful machine learning algorithms may be added.