

Supervised Machine Learning:

Classification

IBM Skills Network

Project Report

**By:**

**Jato Daniel**

Table of Contents

[Main Objective](#_Toc113374188) 2

[Project Description](#_Toc113374189) 2

Data Cleaning [& Feature engineering 5](#_Toc113374190)

[Classifier Models 7](#_Toc113374191)

[Best Performing Model](#_Toc113374192) 8

[Key Findings and Conclusions 9](#_Toc113374193)

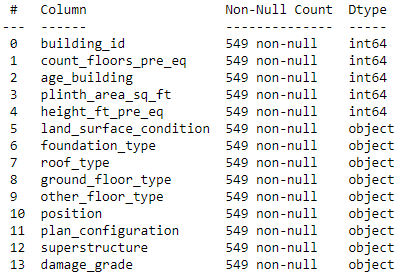
[Suggestions](#_Toc113374194) 10

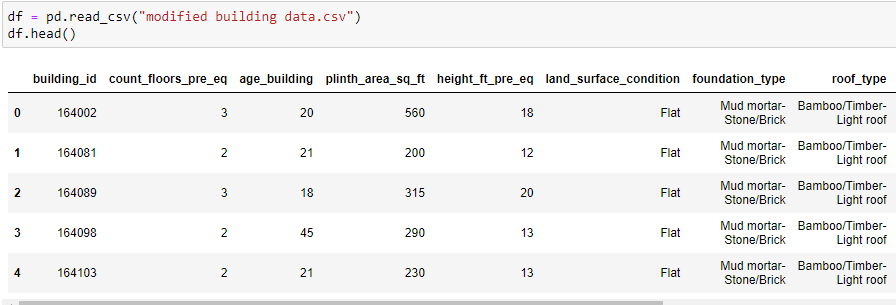
# Main Objective

The main objective of the analysis is to predict if the structure of a build is subject to damage. This has become necessary due to the impact of environmental and climatic changes on the structure of a building. Conventional methods of building damage estimation require a large amount of time and financial resources. For this reason, in recent years, machine learning algorithms that produce faster and more economical results have become the research topics of interest in damage estimation.

# Brief Description

The dataset consists of features that would help us identify and classify building structure that are prone to severe damage. The dataset consists of 550 rows and 14 columns with severe damage column as our target. Shown below is a brief information of the features and data types of the dataset.



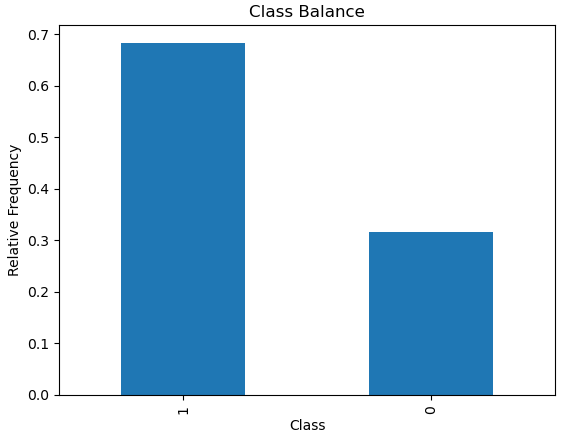


# Data cleaning & Feature engineering

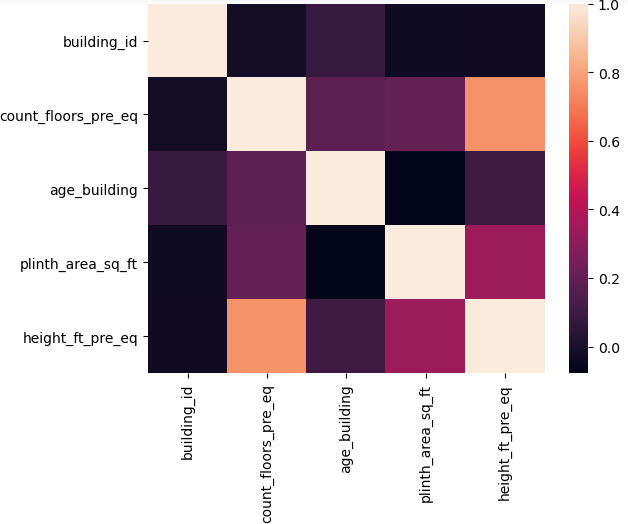
Data wrangling, Exploratory data analysis and feature engineering were performed before training the data.

The dataset contains no missing value. columns with high cardinality such as the building\_id column was dropped as this column is not necessary for our machine learning.

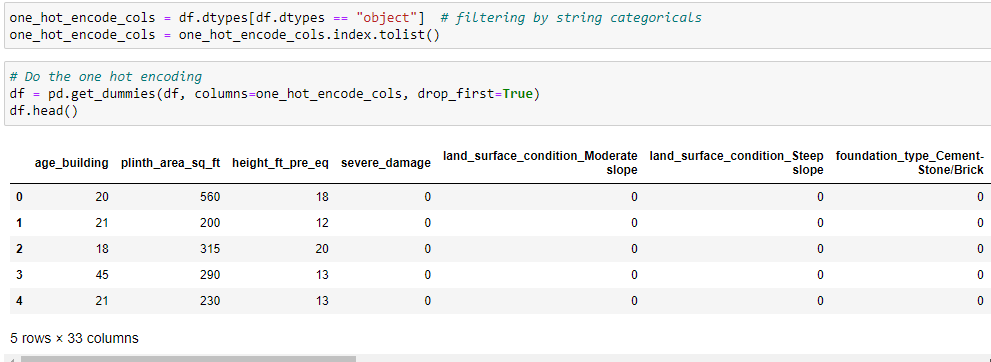
There was imbalance in the dataset as the proportion of severe damages far outweighed the proportion of non-severe damages as shown below.



We checked for multicollinearity in the dataset by performing a heat map plot to show the correlations among the features.

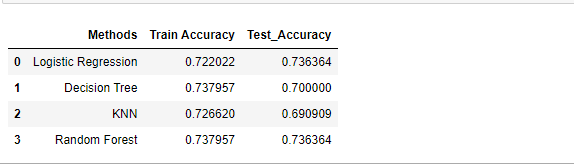


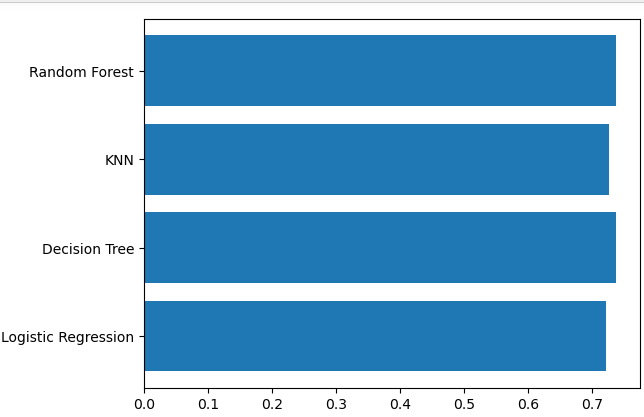
OnehotEncoder was performed in order to convert the categorical columns to numerical columns and also standard scaling was applied to the dataset as shown below.



# Classifier Models

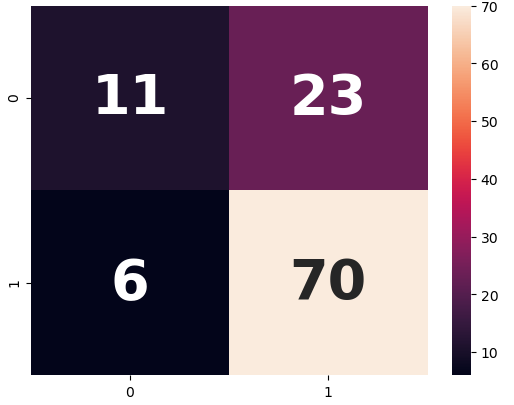
The dataset was trained using logistic regression and three other classifiers such as KNearest neighbor classifier, Decision tree classifer and Random Forest classifier. All models used the same training and test splits and the same cross-validation method. The models were tune using hyper parameters and the best parameters were used for the prediction. Summary of the outcome of the result of the models on both the training and test data is shown below.





# Best Performing Model

Random forest seems to yield the best result when compared to other models in consideration. The model did better at predicting buildings with severe damages but did fairly at predicting building structure with non-severe damages. The confusion matrix of random model classifier is shown below.



# Key Findings and Insights

KNearest Neighbor classifier did poorly on predicting building structure with severe and non-severe when compared with other models under consideration. Random Forest Classifier performs best. Random Forest Classifier had a training accuracy of 73.8% and test accuracy of 73.6%. All the models under consideration did fairly well at predicting building structure with non-severe damages. This is due to the imbalance of the dataset.

# Suggestion

Based on the findings, it is highly recommended that there should be a balanced dataset for accurate predictions. Also, it is discovered that the more the number of rows of the dataset, the more likely that the accuracy of the result. I will also recommend the dataset should be tune with some other hyperparameters for the possibility of a better prediction.