**Project Report Overview**

* **Project** **Title**: House Prices: Advanced Regression Techniques
* **Project Link:** <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>
* **Project Description** : In this challenge, we intend to use various variables to be able and determine the sale price of a house. A training and testing data is provided along with the description of each column or variable used to determine the price of the house.
* **Project Detail :** With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges us to predict the final price of each home.
* **Project Goals :** To predict the sales price for each house. For each Id in the test set, we intend to predict the value of the SalePrice variable.

* **Data Overview**

The data that were given are spreadsheets and textfiles

For example: sample\_submission.csv reviews what an acceptable submission would look like, test.csv is the test data provided, train.csv is the training data, and data\_description.txt describes what the data look

* **Developed Method, Tools, Algorithm for Project’s Requirements**

We used supervised learning method to develop the project.

We have trained possible such linear regression, decision trees, random forest, and feature selection algorithm to automatically or manually select those features which contribute most to our prediction variable or output in which we are interested in. 10-fold cross validation was applied to linear and logistic regression.

The method to turn the string data into numerical data was to first combine both files from the project. The train and test files from the second index to the last index, excluding 'Sale Price' was combined.

The features with missing data were found, and the features that had over 80% of its data missing was removed.

Features that had less than 80% of its data missing was filled in with data that implies the feature does not have that feature in the house (i.e. BsmtCond, basement condition, was marked as no basement, if data wasn't found.)

We then used df = pd.get\_dummies(df) to transform all the data into numerical data.

Matrices were then created for the training data, and the data was split to have a 0.1 test size with a random state of 3.

Random forest classifier, linear regression, logistic regression, and decision tree classifier were used to see the RMSE and accuracy of our methods.

We did not use KNN, due to having too many features.

* **Responsibility**

**Jateni:** Analyzing the data we are going to work on

**Norin:** Train andchoose the models for project.

**Koung:** Finding best features, transforming features into numerical data, training, and evaluating the models.

**Nicholas:** Feature reduction, training, and model evaluation.