Mini-Project Presentation

Presented by Kai Kai
Heng Cheng Jin, Dexter
Georemon Russelraj
(Dataset 2)

Key Question:

How does height of a building affect the damage grade sustained?

Steps:

1. Exploratory analysis
2. Preparation of data
3. Machine learning
4. Evaluation of models
5. Testing best model
6. Using model for prediction

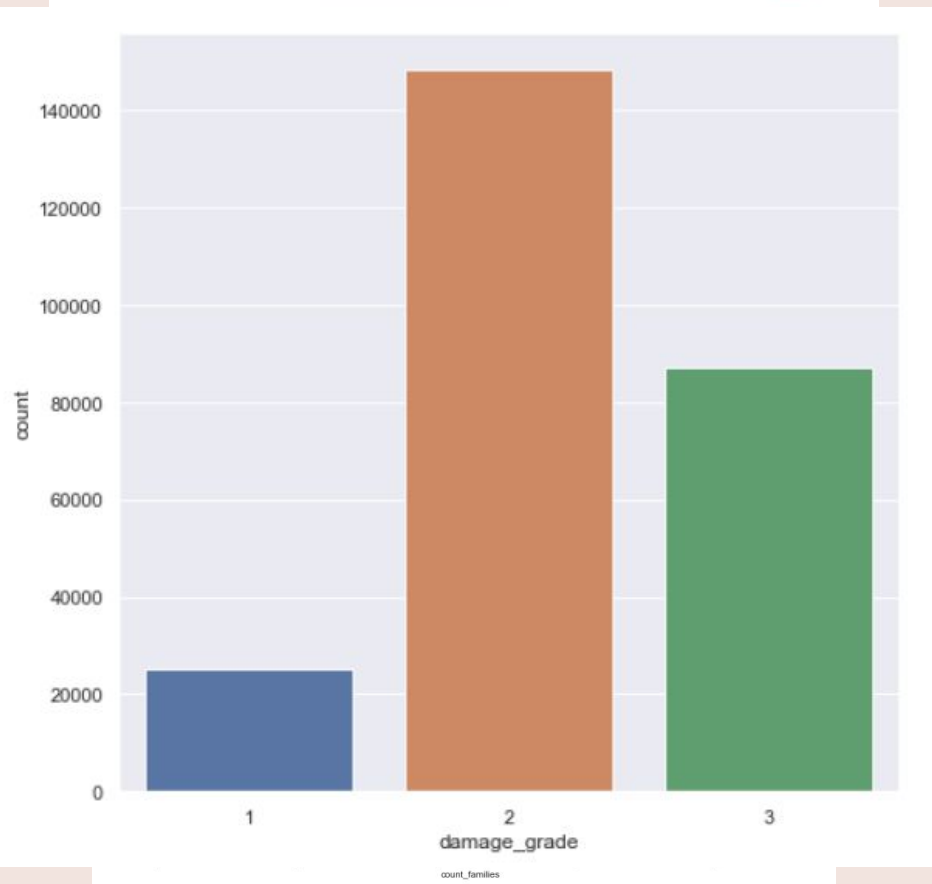# Exploratory data analysis

01 Numerical data distribution.

02 Numeric data vs. damage grade.

03 Categorical data vs. damage grade.

04 Height percentage and damage grade.

Preparation of dataset

Splitting of data

Split data into short and tall buildings

Use of SMOTE

Oversampling technique for imbalance datasets.

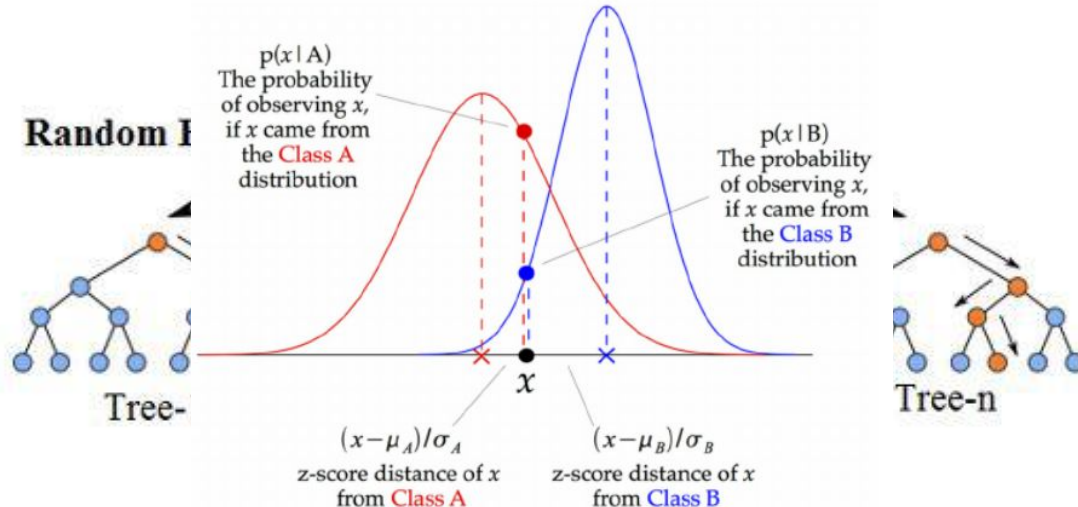# Machine Learning using different algorithms

01
Logistic regression

02
Random forest classifier

03
Gaussian naive bayes

Random F...

$p(x \mid A)$
The probability of observing x, if x came from the Class A distribution

$p(x \mid B)$
The probability of observing x, if x came from the Class B distribution

Random F...

Tree-...

Tree-n

x

$(x - \mu_A)/\sigma_A$
z-score distance of x from Class A

$(x - \mu_B)/\sigma_B$
z-score distance of x from Class B

Evaluation of models

Cross validation of models on training set.

Performance metrics:
- **Accuracy**
- **Precision**
- **Recall**
- **F1 score**

Cross-validated results

```
Logistic Regression:
Training set:
Accuracy: 0.446, Standard Deviation: 0.003

damage_grade:      1            2            3            Average

Precision:         0.478        0.394        0.427        0.433
Recall:            0.634        0.109        0.596        0.446
F1 score:          0.545        0.170        0.497        0.404


Random Forest Classifier:
Training set:
Accuracy: 0.515, Standard Deviation: 0.003

damage_grade:      1            2            3            Average

Precision:         0.615        0.421        0.498        0.511
Recall:            0.652        0.385        0.510        0.516
F1 score:          0.633        0.402        0.504        0.513


Gaussian NB:
Training set:
Accuracy: 0.447, Standard Deviation: 0.001

damage_grade:      1            2            3            Average

Precision:         0.578        0.356        0.400        0.445
Recall:            0.491        0.084        0.766        0.447
F1 score:          0.531        0.136        0.526        0.398
```
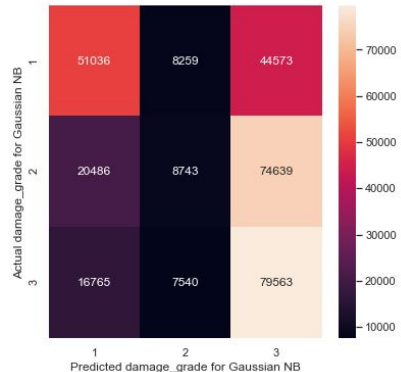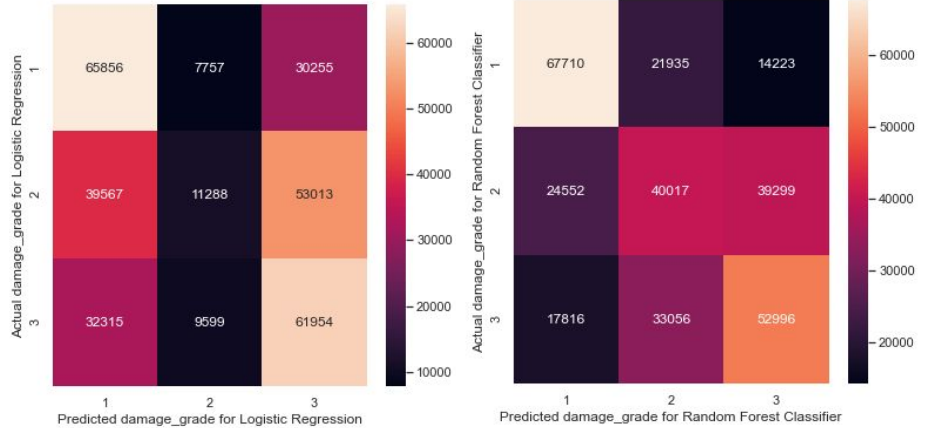
Training and testing of random forest classifier

```
Gaussian NB:
Training set:
Accuracy: 0.442, Standard Deviation: 0.000

damage_grade:     1            2            3            Average

Precision:        0.574        0.346        0.397        0.439
Recall:           0.474        0.073        0.778        0.442
F1 score:         0.519        0.121        0.526        0.389

Test set:
Accuracy: 0.348, Standard Deviation: 0.000

damage_grade:     1            2            3            Average

Precision:        0.207        0.610        0.365        0.394
Recall:           0.477        0.075        0.774        0.442
F1 score:         0.289        0.133        0.496        0.306
```
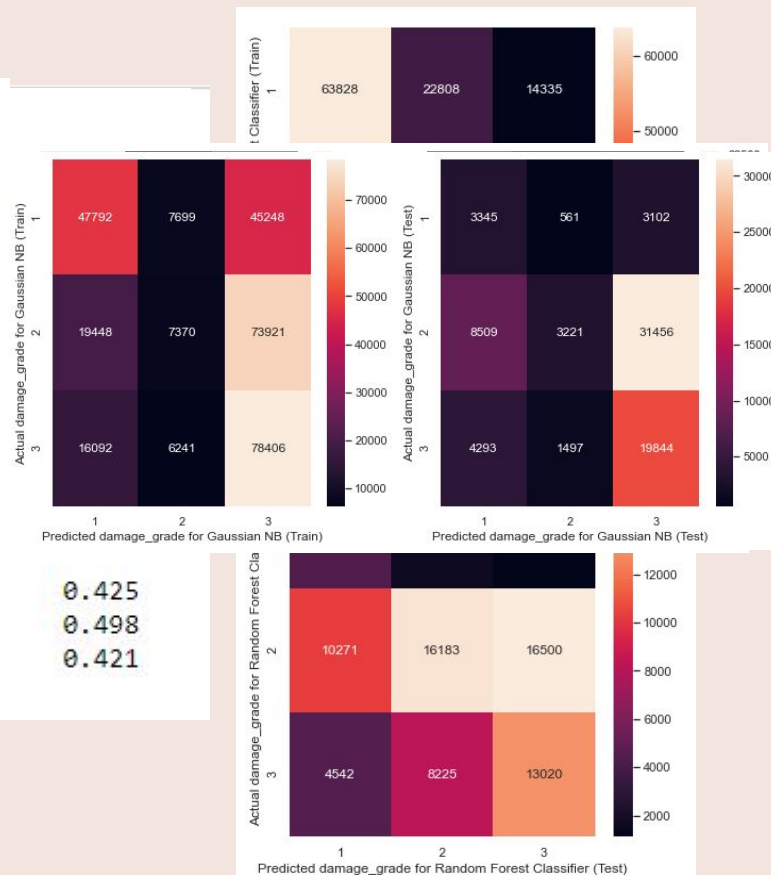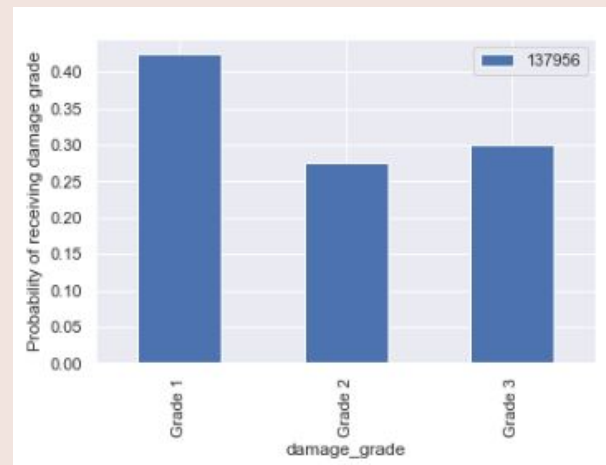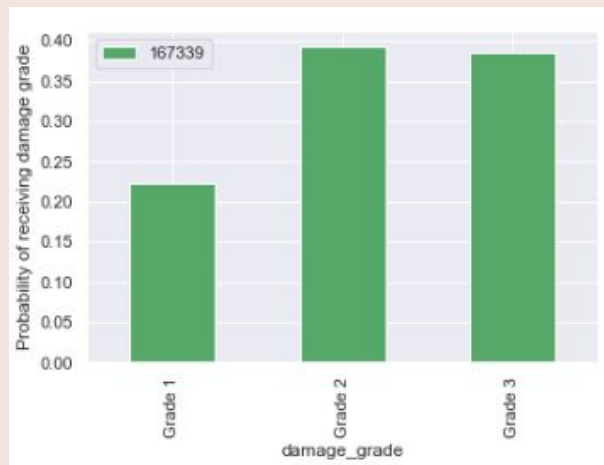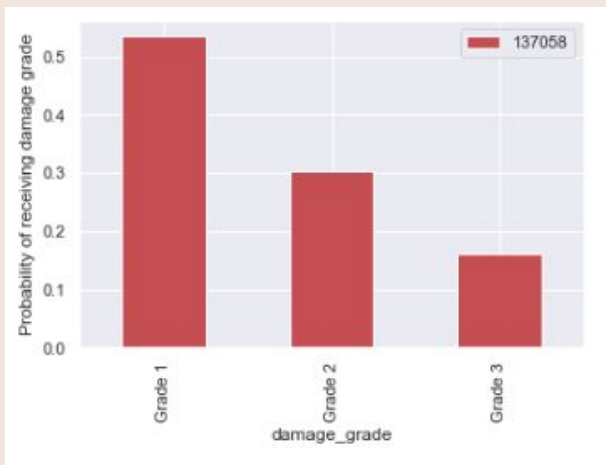
```
Recall:              0.615        0.577        0.585        0.425
F1 score:            0.331        0.469        0.461        0.498
                                                            0.421
```

Predicted Future Damage Grades

	count_floors_pre_eq	age	height_percentage	area_percentage	count_families	damage_grade
195820	3	45	6	8	1	3
24891	3	0	9	6	1	3
82563	2	45	4	13	1	2

Conclusion

01 Shorter buildings are more likely to collapse than taller buildings.

02 Numeric variables are insufficient in determining damage grade.

03 Splitting of data increases accuracy of results.