* Regression:
  + Dataset description:
    - Dataset: Egypt houses prices scrapped from olx
    - Dataset size before cleaning : ~27K row \* 12 column
    - Dataset after cleaning: ~16K row \* 77 columns
    - Train size: 80% of the data
    - Test size : 20% of the data
  + Cleaning and preprocessing:
    - Dropped the unnecessary columns from the data
    - Dropped the outliers:
      * Homes with rooms less than 8m2
      * Homes with 3+bathrooms more than bedrooms
      * Outliers of price per m2 of each type of houses
    - Mapped the categorical data to be numerical
    - Log normalization for the prices
  + Models comparison:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Linear regression (train) | Linear regression (test) | KNN (train) | KNN (test) |
| MAE | 0.28 | 0.28 | 0.18 | 0.23 |
| MSE | 0.12 | 0.12 | 0.06 | 0.1 |
| R2 Score | 0.78 | 0.77 | 0.88 | 0.81 |

* + Linear regression is better as the KNN overfits the data.