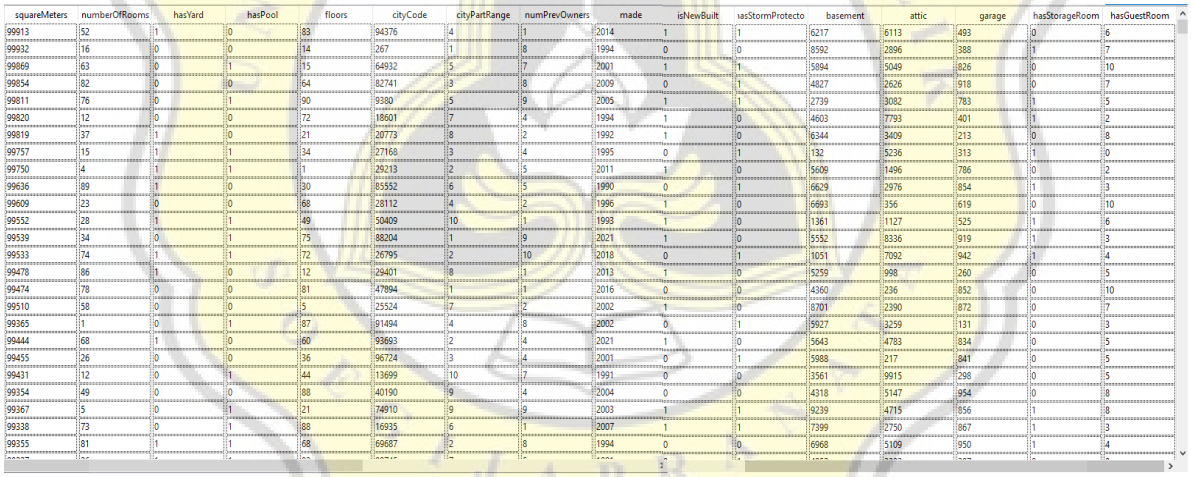
CHAPTER 4
ANALYSIS AND DESIGN

4.1. Collecting data

Dataset was taken from keggle.com with this url https://www.kaggle.com/datasets/mssmartypants/paris-housing-price-prediction, This is a set of data created from imaginary data of house prices in an urban environment, with 10000 total data and 17 attributes attrubut that in the dataset such as, squareMeters, numberOfRooms, hasYard, hasPool ,floors - number of floors, cityCode - zip code, cityPartRange - the higher the range, the more exclusive the neighbourhood is, numPrevOwners - number of prevoious owners, made - year, isNewBuilt, hasStormProtector, basement - basement square meters, attic - attic square meteres, garage - garage size, hasStorageRoom, hasGuestRoom - number of guest rooms, price - price of a house, category - Luxury or Basic.



squareMeters	numberOfRooms	hasYard	hasPool	floors	cityCode	cityPartRange	numPrevOwners	made	isNewBuilt	hasStormProtecto	basement	attic	garage	hasStorageRoom	hasGuestRoom
99913	52	1	0	83	94376	4	1	2014	1	1	6217	6113	493	0	6
99932	16	0	0	14	267	1	8	1994	0	0	8592	2896	388	1	7
99869	63	0	1	15	64932	5	7	2001	1	1	5894	5049	826	0	10
99854	82	0	0	64	82741	3	8	2009	0	1	4827	2626	918	0	7
99811	76	0	1	90	9380	5	9	2006	1	1	2739	3082	783	1	5
99820	12	0	0	72	18601	7	4	1994	1	0	4603	7793	401	1	2
99819	37	1	0	21	20773	8	2	1992	1	0	6344	3409	213	0	8
99757	15	1	1	34	27168	3	4	1995	0	1	132	5236	313	1	0
99750	4	1	1	1	29213	2	5	2011	1	0	5609	1496	786	0	2
99636	89	1	0	30	85552	6	5	1990	0	1	6629	2976	854	1	3
99609	23	0	0	68	28112	4	2	1996	1	0	6693	356	619	0	10
99552	28	1	1	49	50409	10	1	1993	1	0	1361	1127	525	1	6
99539	34	0	1	75	88204	1	9	2021	1	0	5552	8336	919	1	3
99533	74	1	1	72	26795	2	10	2018	0	1	1051	7092	942	1	4
99478	86	1	0	12	29401	8	1	2013	1	0	5259	998	260	0	5
99474	78	0	0	81	47894	1	1	2016	0	0	4360	236	852	0	10
99510	58	0	0	5	25524	7	2	2002	1	0	8701	2390	872	0	7
99365	1	0	1	87	91494	4	8	2002	0	1	5927	3259	131	0	3
99444	68	1	0	60	93693	2	4	2021	1	0	5643	4783	834	0	5
99455	26	0	0	36	96724	3	4	2001	0	1	5988	217	841	0	5
99431	12	0	1	44	13699	10	7	1991	0	0	3561	9915	298	0	5
99354	49	0	0	88	40190	9	4	2004	0	0	4318	5147	954	0	8
99367	5	0	1	21	74910	9	9	2003	1	1	9239	4715	856	1	8
99338	73	0	1	88	16935	6	1	2007	1	1	7399	2750	867	1	3
99355	81	1	1	68	69687	2	8	1994	0	0	6968	5109	950	1	4

Figure 4.1 Data sample

4.2. Linear Regression

Linear regression is an algorithm to make an analysis used to predict the value of a variable based on the value of another variable, and the variable we want to predict is called the dependent variable, A linear regression algorithm in orange using a dataset as input and preprocessing method and outputs are linear regression learning algorithm and also get trained model also a coefficient, linear regression widget in orange constructs a

learner/predictor that learns a linear function from input data, a model can identify the relationship between a predictor (xi) and response (y).

- Add Dataset

In this step in the orange data mining tool, there's a widget to add dataset/import using CSV that the author already downloaded from keggle.com, this step is recommended to configure import settings to make sure data can be imported successfully, such as changing cell delimiter setting match with CSV file.

- Data Splitting

This step have a relation to data that the author already imported, the imported data have 10000 data, and all of those data will divide using a widget data sampler using an orange data mining tool, divided data called data sample and remaining data, for example, if the author uses fixed proportion data at 70%, so data sample or can called training data have 7000 data and 3000 become remaining data or testing data.

- Inputting Model/Algorithm

This step used the Model widget in the orange data mining tool, The author used a linear regression model to get a prediction, with a setting, with 0.1 regularization strength with elastic net mixing 0.50 value both L1 and L2 and also without setting, best output in this widget is without using a setting in Linear regression widget setting.

- Prediction

The last part of using orange data mining is making a prediction, and the prediction output is MSE, RMSE, MAE, R2.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(target - prediction)^2 \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|target - prediction| \quad (3)$$

Indeks	explanation
MSE	Mean Square Error
N	Total Data
RMSE	Root mean Square error
MAE	Mean absolute error
∑	Summation

4.3. Random Forest

Random Forest (Random Decision Forest) is an ensemble learning method for classification, regression, and other tasks entirely by building divergent decision trees until buildup. For classification tasks, the output of a random forest is the classes selected by most trees.

Random forest widget In orange, a widget to construct a set of decision trees. Each tree is constructed from a bootstrap instance of the training data. When building individual trees, an arbitrary subset of the attributes is drawn (the term "random" is used), from which the best attributes are conscripted and separated. The final model is based on a majority vote of individually developed trees within the forest.

- Add Dataset

The first step same as using the linear regression method which is importing a CSV file using a widget in the orange data mining tool, and data that the author gets from downloading from keggle.com, and also imported CSV needs some configuration such as changing the cell delimiter

- Data Splitting

After adding the dataset, the next step is using the data sampler widget in the orange data mining tool that can divide total data become training data and testing data, for an example if there's a 1000 data, using fixed 70% makes training data become 700 and testing data have total 300

- Inputting model/Algorithm

This step uses a model in the orange data mining widget, and that model name is the random forest, with configuration number of trees, replicate training, and also

balance class distribution, this step is to make sure better accuracy with or without configuration to get a better result

- Prediction

The last part of using orange data mining is making a prediction, and the prediction output is MSE, RMSE, MAE, R2.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(target - prediction)^2 \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|target - prediction| \quad (3)$$

Indeks	explanation
MSE	Mean Square Error
N	Total Data
RMSE	Root mean Square error
MAE	Mean absolute error
∑	Summation

4.4. Desain

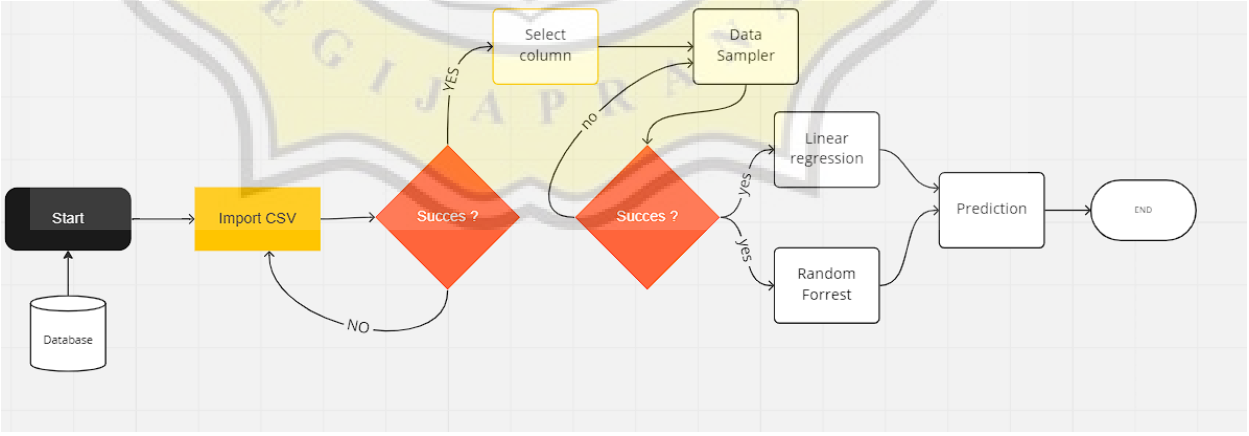


Figure 4.1 Flowchart

The first thing to do in this flowchart, enters dataset into Orange. Enter CSV file if successful next is the select column for attributes and then next is the data sampler if sampler data is a success next is the select method if no goes to the data sampler option again to find which one is an error, in next step is using both method widget and the last is using prediction widget.

4.5. Result

After going through the regression handle between the two algorithms and getting prediction outputs such as MSE, MAE, RMSE, and R2. The author has thought that both algorithms have their advantages and disadvantages in terms of accuracy and error values, in each of them.

4.6. Writing report

The author will make a report discussing the process of making this study, from importing data to making predictions. This study also discusses the comparison between the linear regression algorithm and the random forest algorithm, which one of them has less error and better accuracy in predicting prices.