A Comparison of Clustering Machine Learning Models in Discovering Patterns of Apartments in Riyadh, Saudi Arabia

Manal Anetallah Alsahafi

Introduction

This report aims to investigate the performance of two of the clustering machine learning algorithms to discover the patterns in the apartments of Riyadh city where apartments with similar attributes are placed on one cluster. These attributes are how old is the apartment, how many long the apartment was posted on the website before being deleted, how many beds rooms, living rooms, and bathrooms there are, does the apartment contain facilities or not (kitchen, furnished), the size of the apartment, the price, the neighborhood name and the width of the street.

Data

The dataset that has been used in this work is a public dataset from from Kaggle called "Apartments in Riyadh Saudi Arabia". It has been collected and scraped from AQAR website which is considered the largest online real estate listing company in Kingdom of Saudi Arabia, allows agents and sellers to connect to renters and buyers all over the country [2]. The dataset contains only the data of apartments in Riyadh that have stayed one month from 2022-07-07 to 2022-08-06 on the AQAR website. It has around 6762 observations in it with 15 columns and containe mixual values between categorical and numeric. The sample dataset used illustrated in Table 1.

Table 1. Overview on the dataset

	district	area	age	num_bedrooms	num_livings	num_water_cycles	street_width	IsKetchen	IsFurnished	review	onMarket	IsRent	price
0	حي النظيم	225.0	9.0	3	0.0	2	15.0	1.0	0.0	5.00	17	False	20000
1	حي الفيحاء	130.0	12.0	3	1.0	2	30.0	1.0	0.0	4.33	5	True	25000
2	حي الرمال	200.0	NaN	3	1.0	2	25.0	0.0	0.0	4.67	15	True	22000
3	حي المغيق	120.0	0.0	1	1.0	1	34.0	1.0	0.0	4.17	165	False	38000
4	حي التعاون	60.0	9.0	1	1.0	1	39.0	1.0	0.0	4.42	48	False	25000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6757	حي النرجس	180.0	0.0	3	1.0	2	18.0	1.0	0.0	4.31	33	False	60000
6758	حي غبيرة	90.0	25.0	2	1.0	1	5.0	0.0	0.0	4.82	493	False	9800
6759	حي اليرموك	120.0	5.0	3	1.0	2	20.0	1.0	1.0	5.00	62	False	50000
6760	حي النرجس	200.0	1.0	2	2.0	2	15.0	1.0	0.0	4.54	20	False	40000
6761	حي النرجس	70.0	3.0	1	0.0	1	34.0	0.0	0.0	4.54	192	False	17000

6762 rows × 13 columns

The reason behind clustering the data set is to show how different apartments are similar to others. The first step was cleaning the dataset of missing values by replacing NaN values with zero or dropping its rows. Then, the categorical data converted into numeric and the datatype of int columns change into float. Table 2 shows the dataset after preprocessing steps which decreased to 5882 examples.

Tabel 2: Clean Dataset

	id	district	area	num_bedrooms	num_livings	num_water_cycles	street_width	IsKetchen	IsFurnished	review	onMarket	IsRent	price
0	4596035	حي النظيم	225.0	3	0.0	2	15.0	1.0	0.0	5.00	17	False	20000
1	4599813	حي الفيحاء	130.0	3	1.0	2	30.0	1.0	0.0	4.33	5	True	25000
2	4554519	حي الرمال	200.0	3	1.0	2	25.0	0.0	0.0	4.67	15	True	22000
3	4120004	حي العقيق	120.0	1	1.0	1	34.0	1.0	0.0	4.17	165	False	38000
4	4498954	حي التعاون	60.0	1	1.0	1	39.0	1.0	0.0	4.42	48	False	25000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6757	4538503	حي النرجس	180.0	3	1.0	2	18.0	1.0	0.0	4.31	33	False	60000
6758	3130523	حي عنيزة	90.0	2	1.0	1	5.0	0.0	0.0	4.82	493	False	9800
6759	4453217	حي اليرموك	120.0	3	1.0	2	20.0	1.0	1.0	5.00	62	False	50000
6760	4586116	حي النرجس	200.0	2	2.0	2	15.0	1.0	0.0	4.54	20	False	40000
6761	4025771	حي النرجس	70.0	1	0.0	1	34.0	0.0	0.0	4.54	192	False	17000

6552 rows × 13 columns

Exploratory Data Analysis (EDA)

"EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset" [3]. Different descriptive statistics for numerical data are applied and various graphical representations create a better understanding of the data. The reason for changing all types of the dataset into floats is to use the describe function to get a descriptive statistics summary of our dataset. See figure 1.

	0	1	2	3	4	5	6	7	8	9	10
count	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03	5.015000e+03
mean	7.828013e-17	-1.246815e-16	6.942491e-17	4.108821e-17	-4.958922e-18	-1.501845e-16	7.792592e-18	8.430168e-17	-1.054125e-15	5.242289e-17	1.126384e-16
std	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00	1.000100e+00
min	-2.163026e+00	-1.983241e+00	-1.414156e+00	-2.209080e+00	-1.174120e+00	-1.882342e+00	-2.612224e+00	-2.684014e-01	-2.732023e+00	-1.386093e+00	-1.954473e+00
25%	-7.922174e-01	-6.282065e-01	-5.130132e-01	1.507905e-02	-1.174120e+00	-6.359802e-01	3.828156e-01	-2.684014e-01	-4.888395e-01	-7.322197e-01	-7.891019e-01
50%	2.776818e-01	1.330489e-01	3.881291e-01	1.507905e-02	-8.189000e-02	-1.908509e-01	3.828156e-01	-2.684014e-01	-1.029374e-02	-2.963039e-01	-3.219256e-01
75%	8.460657e-01	5.898022e-01	3.881291e-01	1.507905e-02	1.010340e+00	6.994078e-01	3.828156e-01	-2.684014e-01	5.729339e-01	5.755276e-01	6.124269e-01
max	1.815662e+00	2.599517e+00	3.992698e+00	8.911716e+00	3.194799e+00	2.657977e+00	3.828156e-01	3.725762e+00	1.814162e+00	3.154696e+00	3.181896e+00

Fingure 1: The Descriptive Statistics

For more visualization, we used a boxplot to show the outliers in numeric features like in figure 2 which shows the age, street_width, review, onMarket and price outliers. In addition, the pie plot is used to see the amount of Rental and Not Rental apartments in our dataset as we are interested in it, and figure 3 shows the Not Rental data is more by 10%.

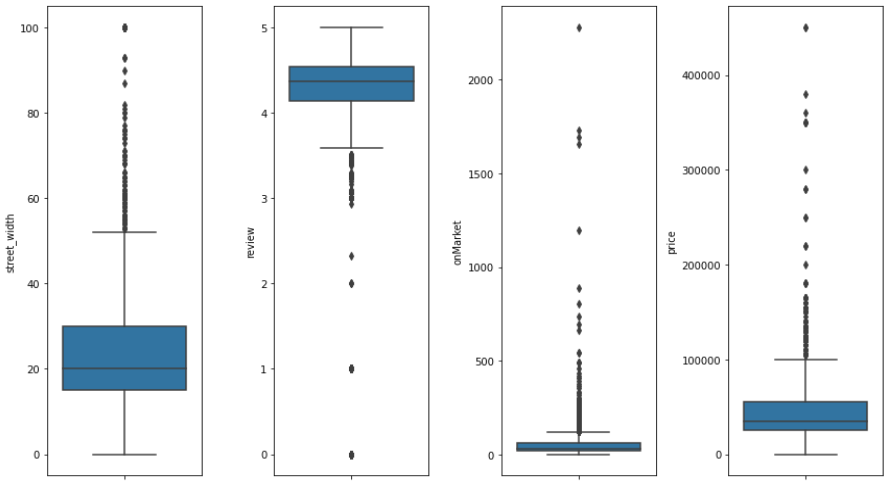
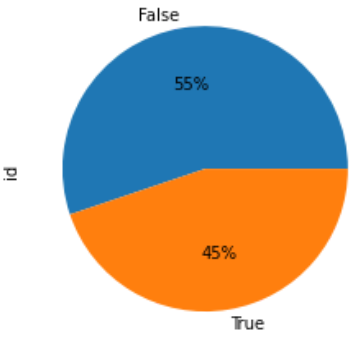


Figure 2: Boxplot of Outliers Features



Figures 3: Pie Plots of Apartments

Feature Engineering

Feature engineering plays a key role in Machine learning and data mining algorithms; the quality of results of those algorithms largely depends on the quality of the available features[4]. The most tasks in feature engineering are: feature transformation, feature generation and extraction, feature selection, automatic feature engineering, and feature analysis and evaluation[4].

In this work, removing the outliers was the first step using IQR on the same numeric features that have outliers. After this process the data of the Rental apartment increased about 2% but still less than Not Rental see figure 4. Then, the entire dataset is scaling using the standard scalar.

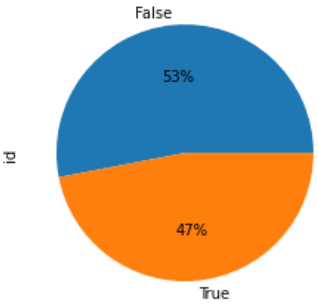


Figure 4 : Distribution of Apartments after removing outliers

Methodology

In this report, we investigated the performance of the most clustering algorithms famous which are the *k*Means model and the Mixture of the Gaussians model to find the optimal number of k and compare it to the original number of classes which is 2. In addition, discussing the use of clustering algorithms for classification problems. To evaluate and compare these models different approach applied.

Results

The first investigation was about the performance of the *k*Means model. In the first step, we pick the number of clusters, k = 2, as we classified the data before into Rental and Not Rental data. The model predicted more labels as rental data and this is not correct as we know the number of not rental is more see figure 5. The optimal number of k as shown in figure 6 is 6 clusters.

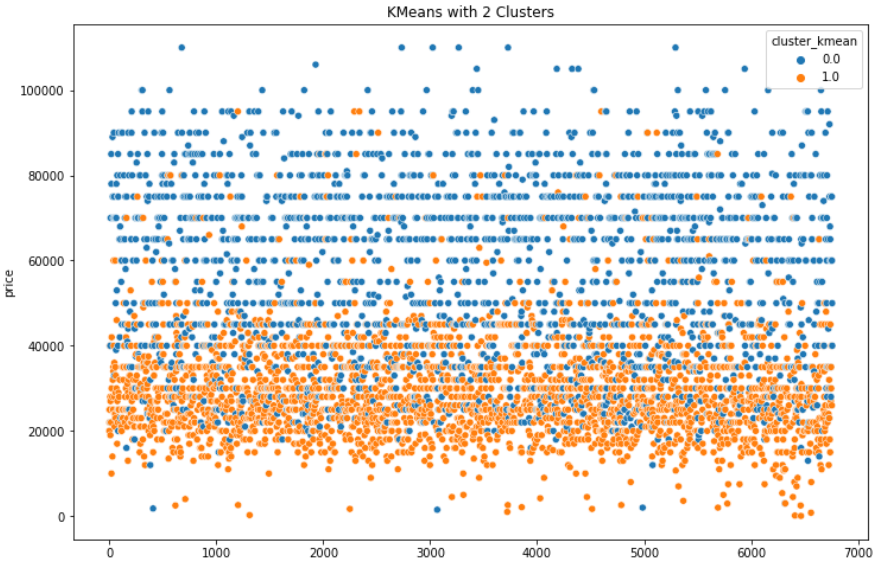


Figure 5: Clustring by Kman model

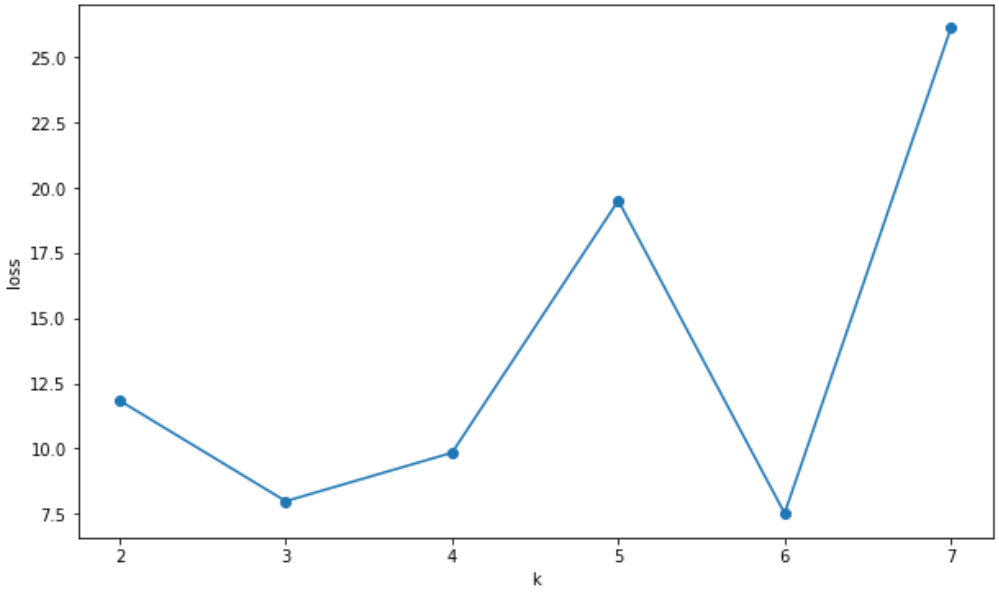


Figure 6: Optimal number of *k* in Kmean model

The second investigation was about the performance of a Mixture of the Gaussians model. The number of clusters, k = 2, as the original number of classes and the original distribution of the dataset can be seen in figure 7.

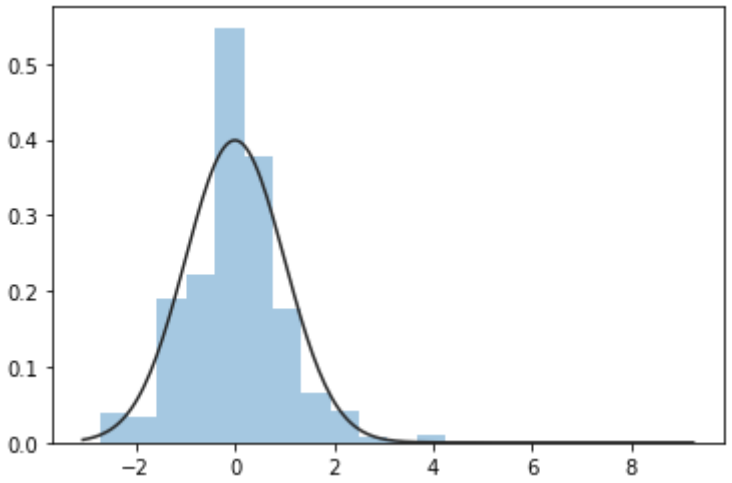


Figure 7: The Original Ddataset Distribution

By experiment, the optimal number of *k* was 2 clusters since the increase of the K number gave a distribution that did not reflect the dataset's original distribution as shown in figure 8, 9.

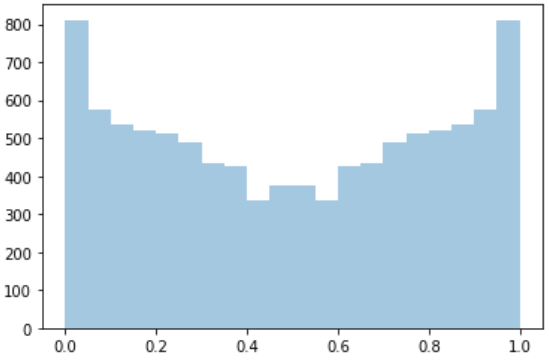


Figure 8: Cluster with 2 K Gaussians models

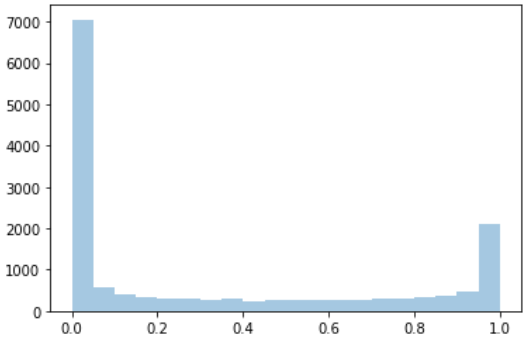


Figure 9: Cluster with 3 K Gaussians models

Discussion and conclusion

This report proves many important insights, in the beginning, the nature of the dataset is very important to determine whether to perform better or worse. The criteria for what defines a good clustering result is also important. A good clustering result was by Mixture of the Gaussian model addressing the weaknesses by generalizing the k-means. The model measures uncertainty in cluster assignment by comparing the distances of each point to all cluster centers, rather than focusing on just distance-from-closet-center. In particular, the non-probabilistic nature of k-means cluster data must be circular, and thus noncircular clusters would be a poor fit in many real-world situations. In addition, using clustering algorithms for classification problems could be very useful if the result of the similarity cluster was good as features that would help to increase the accuracy in classification models.