A Comparison of Binary Classification Machine Learning models in Predicting whether the Apartment will Rent or Not in Riyadh, Saudi Arabia

Manal Anetallah Alsahafi

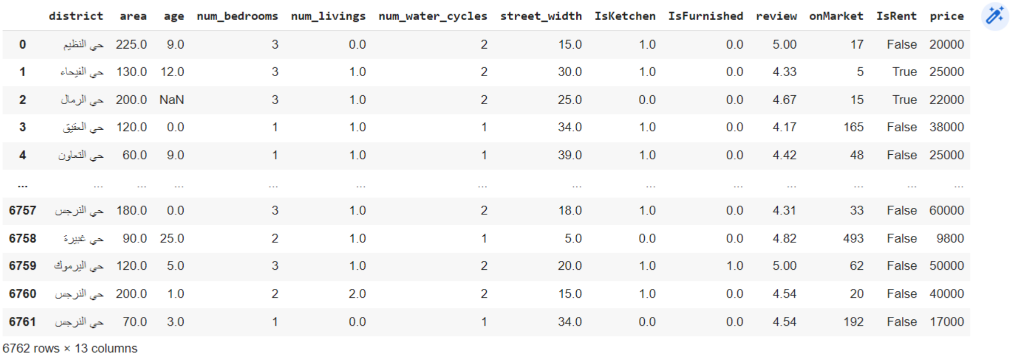
Introduction

The goal of this report is to investigate the performance of different machine learning algorithms in binary classification to predict if the apartment will be rented or not. Riyadh is the capital of Saudi Arabia and the most populous city of the country. It has many branch municipalities and each one containing several districts, amounting to over 130 in total [1]. Therefore, this report aimed to investigate the ability of four classification algorithms where an apartment in Riyadh city will be rented based on many variables which 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. The five models used in our investigation are; a logistic regression in comparing our best model performance with the sklearn model, a linear perceptron to see if it works with our data or not, a Naïve Bayes model performance, a SVM classifier and a Gaussian Discriminant Analysis model in comparing its result with previous models.

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

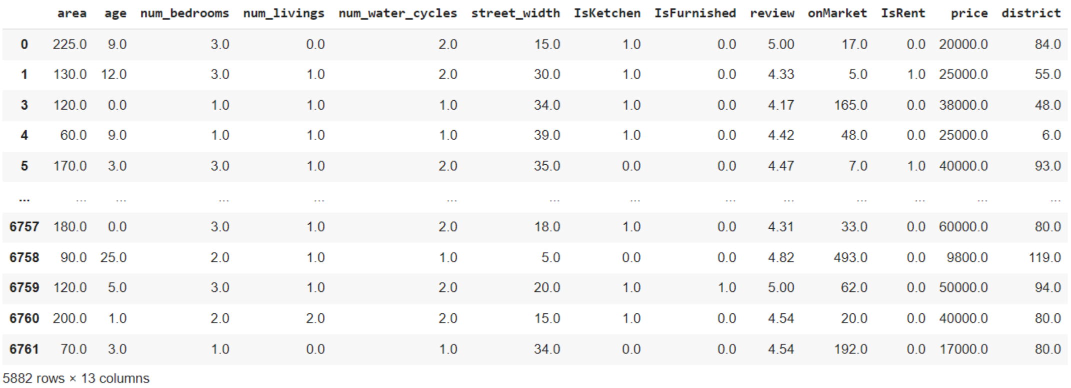


	district	area	age	num_bedrooms	num_livings	num_water_cycles	street_width	IsKetchen	IsFurnished	review	onMarket	IsRent	price
0	حي النظيم	225.0	9.0	3	0.0	2	15.0	1.0	0.0	5.00	17	False	20000
1	حي الفيحاء	130.0	12.0	3	1.0	2	30.0	1.0	0.0	4.33	5	True	25000
2	حي الرمال	200.0	NaN	3	1.0	2	25.0	0.0	0.0	4.67	15	True	22000
3	حي العقيق	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 collecting the dataset is to show how long apartments stayed on the market. 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



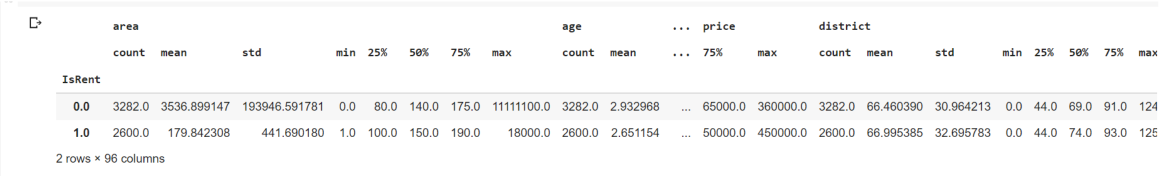
	area	age	num_bedrooms	num_livings	num_water_cycles	street_width	IsKetchen	IsFurnished	review	onMarket	IsRent	price	district
0	225.0	9.0	3.0	0.0	2.0	15.0	1.0	0.0	5.00	17.0	0.0	20000.0	84.0
1	130.0	12.0	3.0	1.0	2.0	30.0	1.0	0.0	4.33	5.0	1.0	25000.0	55.0
3	120.0	0.0	1.0	1.0	1.0	34.0	1.0	0.0	4.17	165.0	0.0	38000.0	48.0
4	60.0	9.0	1.0	1.0	1.0	39.0	1.0	0.0	4.42	48.0	0.0	25000.0	6.0
5	170.0	3.0	3.0	1.0	2.0	35.0	0.0	0.0	4.47	7.0	1.0	40000.0	93.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6757	180.0	0.0	3.0	1.0	2.0	18.0	1.0	0.0	4.31	33.0	0.0	60000.0	80.0
6758	90.0	25.0	2.0	1.0	1.0	5.0	0.0	0.0	4.82	493.0	0.0	9800.0	119.0
6759	120.0	5.0	3.0	1.0	2.0	20.0	1.0	1.0	5.00	62.0	0.0	50000.0	94.0
6760	200.0	1.0	2.0	2.0	2.0	15.0	1.0	0.0	4.54	20.0	0.0	40000.0	80.0
6761	70.0	3.0	1.0	0.0	1.0	34.0	0.0	0.0	4.54	192.0	0.0	17000.0	80.0

5882 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.

Fingure 1: The Descriptive Statistics

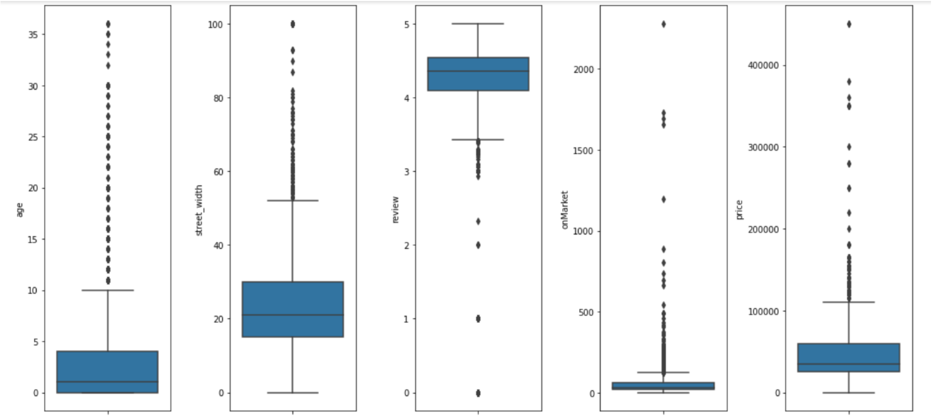


	area								age		...	price		district							
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
IsRent																					
0.0	3282.0	3536.899147	193946.591781	0.0	80.0	140.0	175.0	11111100.0	3282.0	2.932968	...	65000.0	360000.0	3282.0	66.460390	30.964213	0.0	44.0	69.0	91.0	124
1.0	2600.0	179.842308	441.690180	1.0	100.0	150.0	190.0	18000.0	2600.0	2.651154	...	50000.0	450000.0	2600.0	66.995385	32.695783	0.0	44.0	74.0	93.0	125

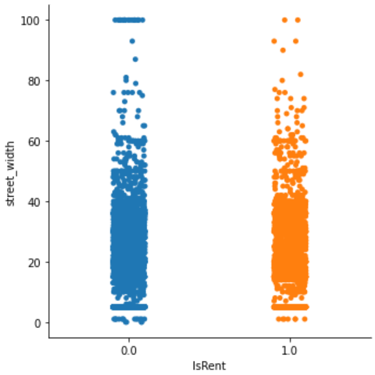
2 rows × 96 columns

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 catplot used for each of these features with respect to IsRent as we are interested in it and figure 3 shows the example of this kind of graphical representation.

Figure 2: boxplot



Figures 3: Example of catplots

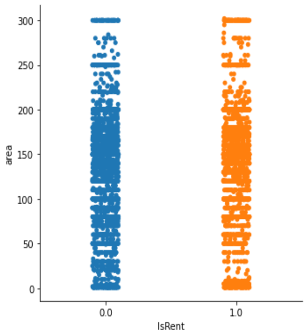


Feature Engineering

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].

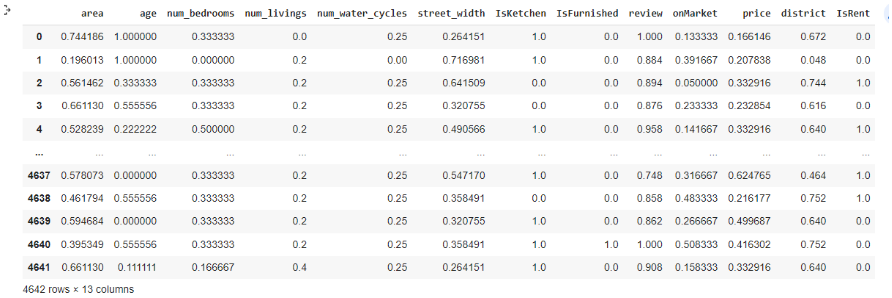
In this work, removing the outliers was the first step using IQR on the same numeric features that have outliers. Figure4, show the age as example of that.

Figure4 : Age feature after removing outliers



Then, scalling dataset after we split them by using a ratio of 70:30, 70% training data, and 30% test data see figure5.

Figure 5: The dataset after Scaling

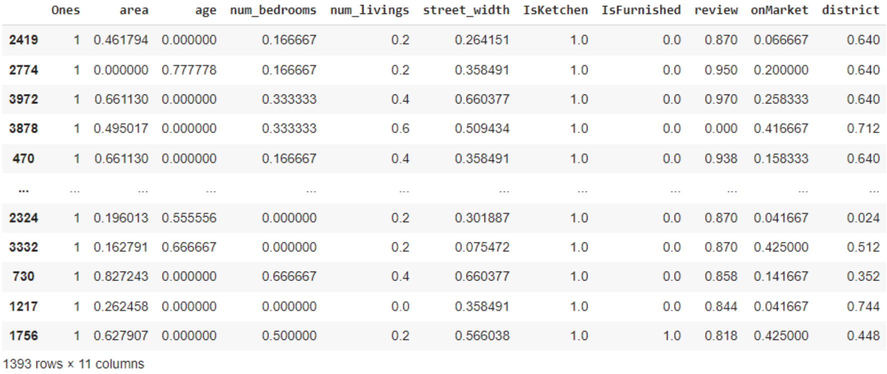


	area	age	num_bedrooms	num_livings	num_water_cycles	street_width	IsKetchen	IsFurnished	review	onMarket	price	district	IsRent
0	0.744186	1.000000	0.333333	0.0	0.25	0.264151	1.0	0.0	1.000	0.133333	0.166146	0.672	0.0
1	0.196013	1.000000	0.000000	0.2	0.00	0.716981	1.0	0.0	0.884	0.391667	0.207838	0.048	0.0
2	0.561462	0.333333	0.333333	0.2	0.25	0.641509	0.0	0.0	0.894	0.050000	0.332916	0.744	1.0
3	0.661130	0.555556	0.333333	0.2	0.25	0.320755	0.0	0.0	0.876	0.233333	0.232854	0.616	0.0
4	0.528239	0.222222	0.500000	0.2	0.25	0.490566	1.0	0.0	0.958	0.141667	0.332916	0.640	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4637	0.578073	0.000000	0.333333	0.2	0.25	0.547170	1.0	0.0	0.748	0.316667	0.624765	0.464	1.0
4638	0.461794	0.555556	0.333333	0.2	0.25	0.358491	0.0	0.0	0.858	0.483333	0.216177	0.752	1.0
4639	0.594684	0.000000	0.333333	0.2	0.25	0.320755	1.0	0.0	0.862	0.266667	0.499687	0.640	0.0
4640	0.395349	0.555556	0.333333	0.2	0.25	0.358491	1.0	1.0	1.000	0.508333	0.416302	0.752	0.0
4641	0.661130	0.111111	0.166667	0.4	0.25	0.264151	1.0	0.0	0.908	0.158333	0.332916	0.640	0.0

4642 rows × 13 columns

To fast the learning of our model and improve the accuracy we used Pearson's Correlation as a selection features method as it deals with the linear dependency between two continuous variables X and Y. In the pairs of features that are strongly correlated with each other, one of them should be removed to help a machine learning model to be more generalized and interpretable [6]. We chose the value of the correlation between variables greater than 0.6 to be removed. The number of bathrooms in the apartment and the price have the correlation columns see Figure 6 after we remove it from the dataset.

Figure 6: The dataset after drop the correlation columns



	Ones	area	age	num_bedrooms	num_livings	street_width	IsKetchen	IsFurnished	review	onMarket	district
2419	1	0.461794	0.000000	0.166667	0.2	0.264151	1.0	0.0	0.870	0.066667	0.640
2774	1	0.000000	0.777778	0.166667	0.2	0.358491	1.0	0.0	0.950	0.200000	0.640
3972	1	0.661130	0.000000	0.333333	0.4	0.660377	1.0	0.0	0.970	0.258333	0.640
3878	1	0.495017	0.000000	0.333333	0.6	0.509434	1.0	0.0	0.000	0.416667	0.712
470	1	0.661130	0.000000	0.166667	0.4	0.358491	1.0	0.0	0.938	0.158333	0.640
...	...	...	...	...	...	...	...	...	...	...	...
2324	1	0.196013	0.555556	0.000000	0.2	0.301887	1.0	0.0	0.870	0.041667	0.024
3332	1	0.162791	0.666667	0.000000	0.2	0.075472	1.0	0.0	0.870	0.425000	0.512
730	1	0.827243	0.000000	0.666667	0.4	0.660377	1.0	0.0	0.858	0.141667	0.352
1217	1	0.262458	0.000000	0.000000	0.0	0.358491	1.0	0.0	0.844	0.041667	0.744
1756	1	0.627907	0.000000	0.500000	0.2	0.566038	1.0	1.0	0.818	0.425000	0.448

1393 rows × 11 columns

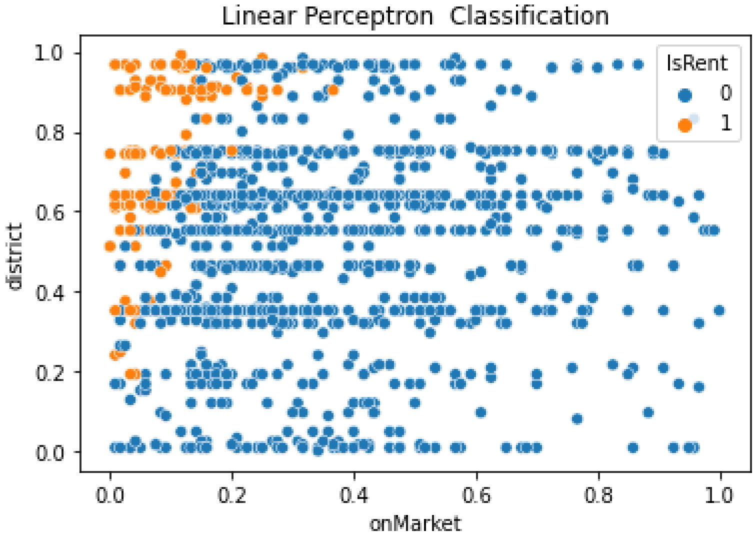
Methodology

In this report, we investigated the steps taken to achieve the desired output by implementing six different experiments with different classification models which are Logistic Regression, SVM, linear perceptron, Naïve Bayes and Gaussian Discriminant Analysis model. To evaluate and compare these models f1-score was chosen.

Results

The first investigation was about the performance of the Regularized Logistic Regression model. The final hyperparameter was chosen for lambda in regularization term was 60 where the iteration value and learning rate were fixed for all experiments at (0.1, 10000) respectively. The second investigation was about comparing our model with sklearn Logistic Regression model and the thierd one was about the performance of a linear perceptron model with our data. The accuracy of this model was the worst since the dataset is not linearly separable see figure 7.

Figure7: Classification by Linear Perceptron model



The fourth investigation was about the performance of a Naïve Bayes model where we get the same accuracy of our model. The fifth one was about comparing our model with SVM classifier and the last investigation was about comparing Gaussian Discriminant Analysis GDA result with previous models. All these comparing you can see in table 1.

Table 1 : The comparison of the Models

	Logistic Regression	Logistic Regression Scklearn	linear perceptron	Naïve Bayes	SVM	GDA
accuracy	0.640	0.647	0.360	0.640	0.643	0.514

Discussion and conclusion

This report proves many important insights, in the beginning, the nature of the dataset is very important if it is linear or nonlinear to choose the best model. After all this effort in preprocessing steps, feature engineering, and improvements in technique the result of our model compared with the other model was very good.