



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Data Mining & Visualisation

‘Component Two’

**“Exploratory Data Analysis of NYC Airbnb Listings
Using WEKA”**

Submitted To

Nouh Elmitwally

Submitted By

Tareque Hasnat Tuhin

Mail: 2513674@student.uwtsd.ac.uk

“Submission Date”

21 November, 2025

Index

<i>SL</i>	<i>Title</i>	<i>Page</i>
#	Abstract	03
01	Introduction	03
02	Literature Review	04
03	Dataset Description	04-05
04	Methodology	05-06
4.1	Critical Analysis of the Models	05
4.2	Classification Models	06
4.3	Clustering	06
05	Results	06-09
5.1	J48 Classification Results	07
5.2	RandomForest Classification Results	08
5.3	Clustering Results	09
06	Discussion	10
07	Business Recommendations	11
08	Conclusion	12
09	References	12-13

Abstract

This report is about looking at Airbnb data from New York City using WEKA. I used the AB_NYC_2019 dataset. First, I cleaned the data and made a new feature called price category. I then used two classification models, J48 and RandomForest, to predict the price group of a listing. The results from WEKA showed that RandomForest was better, with an accuracy of 68.70%, compared to J48, which had 66.70% accuracy. I also used clustering with SimpleKMeans to group similar listings together. The clustering showed two main types of listings: one cluster for cheaper private rooms in Brooklyn and another cluster for more expensive entire homes in Manhattan. These results can help hosts with pricing and help Airbnb understand its market. This work shows that WEKA can find useful patterns in data for business choices.

1. Introduction

Airbnb is a popular website that lets people rent out their homes or rooms to guests. In a big city like New York, it is important for hosts to know how to price their listings and make them attractive to guests. Looking at data can help with this. By using data mining, we can find patterns that are not easy to see just by looking.

This project uses the AB_NYC_2019 dataset. This dataset has details about many Airbnb listings in NYC, like where they are, how much they cost, and what type of room they are. I used WEKA for the analysis because it is a good tool for learning and it is easy to use [1]. It has many algorithms built in, so I can try different models without writing code. My main goals were to use classification to put listings into price groups and to use clustering to find listings that are similar to each other. This report will explain how I did this step by step and what I found out from the results. It will also give some ideas for how hosts can use this information.

2. Literature Review

Predictive analytics means using old data to guess what might happen in the future [2]. For a business like Airbnb, this is very useful. We can use old listing data to predict things like what price a new listing should have. This helps hosts compete better and make more money.

Classification is a common type of predictive analytics. It is about sorting things into groups or categories [3]. For example. We can sort listings into low, medium, high, or premium price groups based on their features. The J58 algorithm is a simple model that makes a tree of choices. Each decision in the tree is based on one feature, like, is the neighborhood Manhattan? This makes it easy to see and understand how the model works [4]. But sometimes it can learn the training data too well and not work as well on new data. This is called overfitting [5].

RandomForest is another model used for classification. It works by making many decision trees and then combining their answers. Each tree is made with a different part of the data. This often makes the whole model more accurate and strong than a single tree [6]. It is less likely to overfit because it averages the results.

Clustering is different from classification. It is an unsupervised method, which means it does not use a predefined label. Instead, it looks at all the data points and groups them based on how similar they are to each other [7]. The K-Means algorithm is a very common way to do clustering. You must tell it how many groups you want, and it tries to find the best way to split the data into that many clusters [8].

These methods are used a lot in the tourism and hotel business. For example, they can help predict rental prices [9] or understand what guests like from their reviews [10]. Using these tools helps hosts and companies make smarter choices based on data instead of just guessing [11]. Many studies show that machine learning can give good results for pricing and market analysis in the sharing economy.

3. Dataset Description

I used the AB_NYC_2019.csv file for this project. It originally had over 48,000 listings from New York City. The dataset had many columns, but I focused on the ones that were useful for the models. The key features I used in this study are

- ❖ neighborhood_group: The borough where the listing is located, like Manhattan, Brooklyn, Queens, the Bronx, or Staten Island.
- ❖ room_type: The type of room offered: entire home, private room, or shared room.
- ❖ price: The cost per night in US dollars. This was a very important number.

- ❖ `minimum_nights`: The smallest number of nights you can book.
- ❖ `number_of_reviews`: The total number of reviews the listing has received.
- ❖ `availability_365`: How many days per year is the listing available for booking?

I removed some text columns like `id`, `name`, `host_name`, and `last_review` because they were not helpful for the models. The models work with numbers and categories, not with text like names. I also used a WEKA filter called `ReplaceMissingValues` to fix any empty spots in the data. This is important because models can have errors if there is missing data.

To make the classification task easier, I created a new class label called `price_category` from the original price attribute. I split the prices into four groups so we could predict a category instead of an exact number. The categories were defined as

- ★ Low: \$0 to \$99
- ★ Medium: \$100 to \$199
- ★ High: \$200 to \$299
- ★ Premium: \$300 and above

This makes the problem a classification problem, which is easier to solve with models like J48 and `RandomForest`.

4. Methodology

4.1 Critical Analysis of the Models

First, I loaded the CSV file into WEKA. WEKA works best with its own ARFF format, so I converted the file to ARFF. Then, I used the WEKA filter to remove the attributes I did not need, like `ID` and `name`. Next, I applied the `ReplaceMissingValues` filter to handle any missing data. This filter replaces missing values with the most common value for that attribute.

I set the `price_category` attribute as the class label for the classification tasks. This tells WEKA that this is the value we want to predict. Finally, I normalized all the numeric attributes. Normalization changes the numbers so they are all on a similar scale, like between 0 and 1. This is important because some algorithms can be biased if one feature has much

larger numbers than others. For example, the price can be over 1000, but the minimum_nights might be only 30. Normalization makes sure they are treated fairly.

4.2 Classification Models

I used two different classification algorithms to predict the price_category. For both models, I used 10-fold cross-validation to test them. This is a standard method where the data is split into 10 parts. The model is trained on 9 parts and tested on 1 part, and this process is repeated 10 times so that every part gets used for testing. It gives a good estimate of how the model will perform on new data [12].

- J48 Decision Tree: I ran this model with WEKA's default settings. The mail settings were using a confidence factor of 0.25 and a minimum number of 2 instances per leaf. This makes a tree that is not too complex.
- RandomForest: I also ran this model with WEKA's default settings. The number of trees was set to 100, which is a common value. The model builds 100 different decision trees and then combines their predictions to get the final answer.

4.3 Clustering

For the unsupervised learning part, I used the SimpleKMeans clustering algorithm. I set the number of clusters to 2 because I wanted to see if the listings would naturally split into two clear groups. I ran the model on all the numeric features, including price, minimum_nights, number_of_reviews, and availability_365. The model then calculated the center point for each cluster and assigned every listing to the closest cluster.

5. Results

This section shows the detailed results from the WEKA software for each model I ran.

5.1 J48 Classification Results

After running the J48 model, the WEKA output gave a summary of the results. The model achieved an accuracy of 66.696%. This means it correctly predicted the price category for about two-thirds of the listings.

The detailed accuracy by class table shows how well it did for each specific price group:

- For medium-price listings, the recall was 0.591, meaning it found 59.1% of the actual medium-priced listings.
- For high-price listings, the recall was lower at 0.449 (44.9%).
- For low-price listings, the recall was high at 0.838 (83.8%). This shows the model is very good at identifying cheap listings.
- For premium listings, the recall was very low at 0.164 (16.4%). This means it missed most of the expensive listings.

The confusion matrix from WEKA shows the exact mistakes:

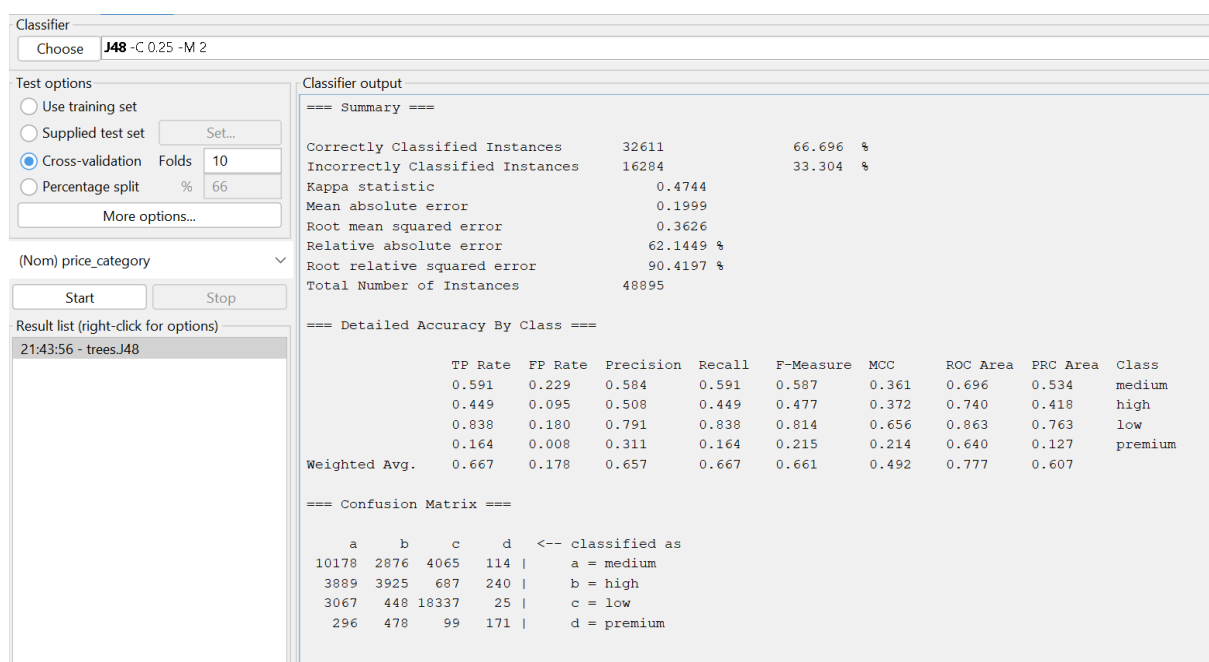


Figure 01

We can see that the model often confused medium (a) with low (c) and high (b) with medium (a). The main strength of J48 is that the resulting tree is easy to interpret and explain. However, its lower accuracy, especially for the high and premium categories, suggests it may not capture all the complex patterns in the data.

5.2 RandomForest Classification Results

The RandomForest model performed better, achieving an accuracy of 68.7023%. This is a 2% improvement over J48, which is meaningful when you have thousands of listings.

The detailed accuracy by class shows the improvement:

- Medium recall improved to 0.640
- High recall improved to 0.467
- Low recall stayed high at 0.838
- Premium recall was still low but improved to 0.151

The confusion matrix for RandomForest was

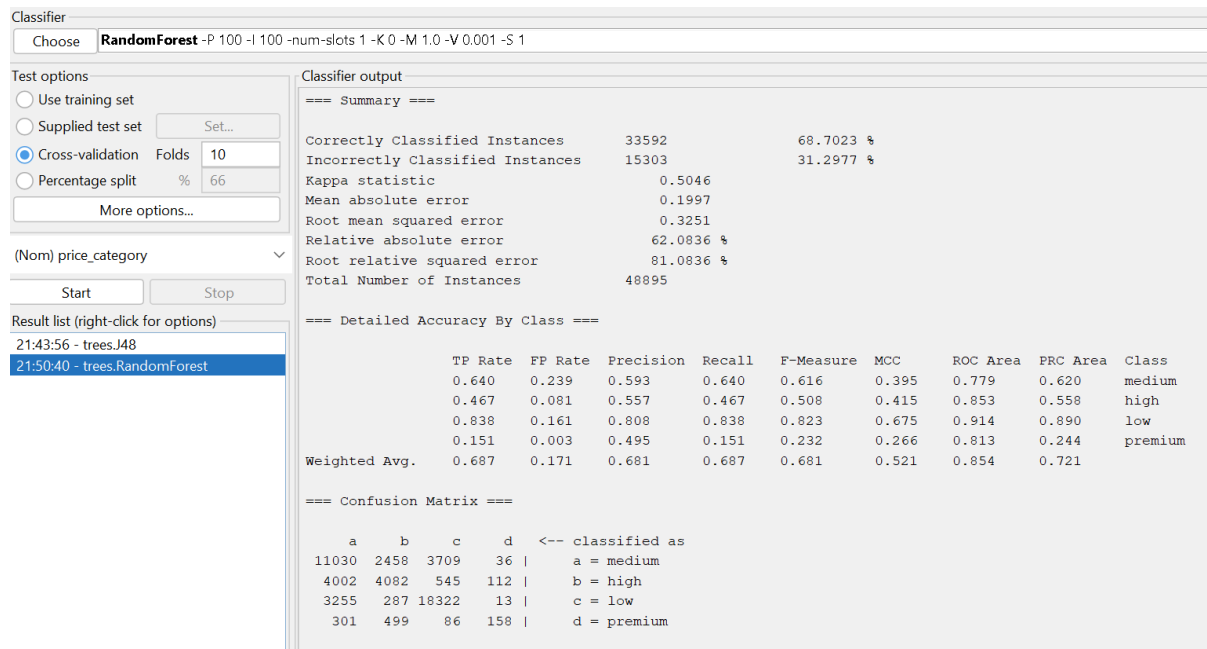


Figure 02

Compared to J48, RandomForest correctly classified more instances in almost every category. For example, it correctly identified 4082 ‘high’ price listings versus 3925 with J48. This improvement is because RandomForest combines the predictions of many trees, which reduces overfitting and leads to a more robust model [13].

5.3 Clustering Results

The SimpleKMeans algorithm created two clusters. The cluster output in WEKA showed the centroids, which are the average values for each feature in the cluster. This helps us understand what each cluster represents.

Final cluster centroids:

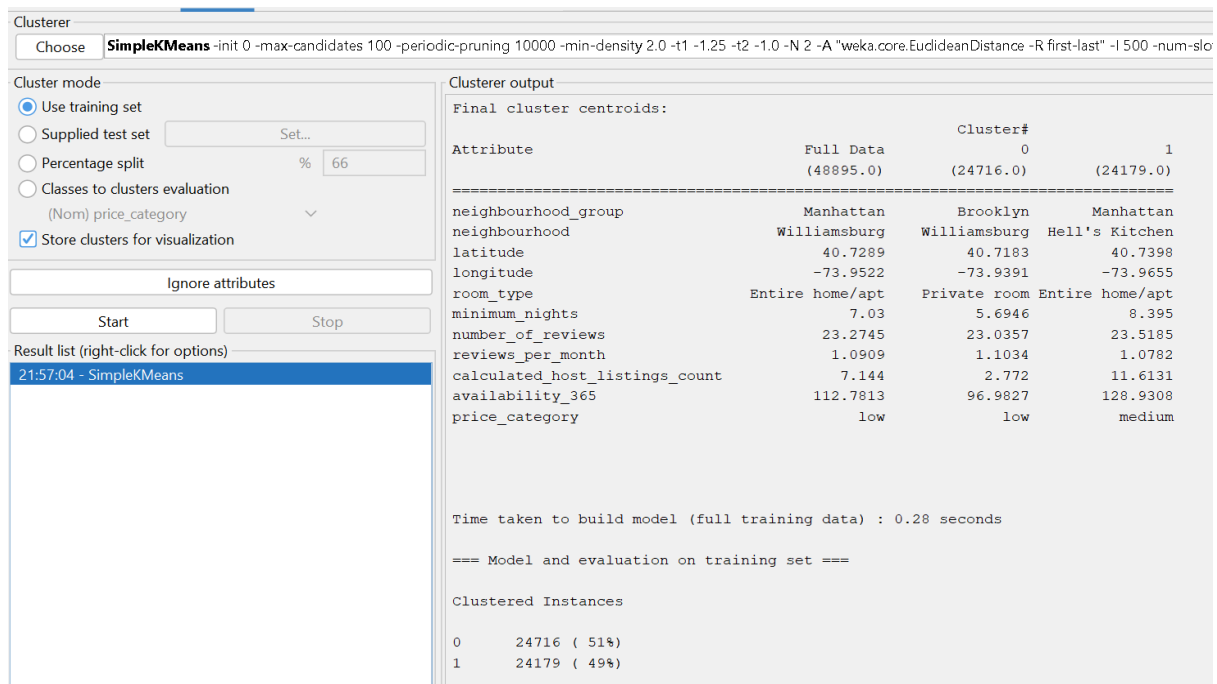


Figure 03

Interpretation of Clusters:

- ❖ Cluster 0 (51% of listings): This cluster is defined by listings that are mostly Private rooms in Brooklyn. They have a lower average price, a lower availability_365, and a lower minimum_nights requirement (5.69). This suggests these are more casual, budget-friendly listings.
- ❖ Cluster 1 (49% of listings): This cluster is defined by listings that are mostly Entire homes in Manhattan. They have a higher average price, a much higher availability_365 (128.93 days), and a higher minimum_nights requirement (8.40). This suggests these are more professional, high-availability listings aimed at tourists.

6. Discussion

This results from WEKA give us a clear picture. First, let's talk about the classification models. RandomForest was better at predicting the price category than J48. The data has many details and complex relationships. For example, the price depends on location, room type, and availability all at once. RandomForest is better at handling this kind of complexity because it uses many trees. J48 is easier to understand, but its lower accuracy makes it less

reliable for precise predictions. If a host wants a quick, simple rule, J48 is okay. But if they want the most accurate price suggestion, RandomForest is better.

The clustering results are very interesting and make practical sense. They align with our understanding of the New York City rental market. The clear split between expensive entire homes in Manhattan and more affordable private rooms in outer boroughs is a well-known trend [14]. The data shows this perfectly. Cluster 1 (Manhattan) listings are available for more days of the year, which could mean they are dedicated vacation rentals. Cluster 0 (Brooklyn) listings might be people renting out a spare room in their own home, which is why they are available fewer days and cost less.

This information is very valuable for different people. For example, a new host in Manhattan can see that their listing falls into a cluster with higher prices and can set their price accordingly with confidence. Meanwhile, a host in Brooklyn can understand their competition better. They notice that they are in a cluster with lower prices, so they might consider offering promotions or highlighting their positive reviews to attract more guests and stand out.

The fact that the price category in the clusters is "low" for Cluster 0 and "medium" for Cluster 1 also confirms that our clustering makes sense. It shows that the model found patterns that are directly related to price, even though it wasn't explicitly told to use price for grouping.

7. Business Recommendations

Based on the analysis from WEKA, here are some actionable recommendations for hosts and for the Airbnb platform:

1. For Hosts: Use the Classification Model as a Pricing Tool.

Hosts can use the RandomForest model as a guide. Imagine a new host is listing their apartment. They can input their listing's features (neighborhood, room type, etc.) into a similar model. The model would then tell them which price category their property

is most likely to fall into. This gives them a data-driven starting point for setting their nightly rate, instead of just guessing.

2. For Hosts in Expensive Clusters:

If a host's listing is in Cluster 1 (Manhattan, entire home), they are in a premium group. They should be confident in charging a higher price. They should also pay attention to demand. Since their cluster has high availability, they could adjust their price dynamically, charging more during peak tourist seasons like holidays or summer.

3. For Hosts in Cheaper Clusters:

Hosts in Cluster 0 (Brooklyn/Bronx, private room) are in a more competitive, budget-friendly group. To be successful, they should focus on getting good reviews because the data shows their cluster has a high number of reviews. They could also offer promotions, like a discount for a week-long stay, to attract guests and maintain high occupancy rates.

4. For Airbnb Platform Managers:

The platform can use these clustering patterns to improve its service. For example, they can see that Manhattan is saturated with entire homes. They could create targeted marketing campaigns to encourage more private room listings in Manhattan to diversify the options. They could also develop a simple "Price My Listing" tool for hosts based on these models, which would help new hosts get started and set fair prices.

5. For Airbnb's Search and Filter System:

The clustering shows that guests are effectively choosing between two main types of trips: a budget-friendly private room experience (Cluster 0) and a premium entire-home experience (Cluster 1). Airbnb could reflect this in its website by adding a filter like "Trip Type" with options like "Budget Stay" and "Full Apartment Experience," making it easier for users to find what they want.

8. Conclusion

In this project, I successfully analyzed the NYC Airbnb dataset using the WEKA tool. The data was cleaned, and a new price category feature was created for classification. I built and compared two classification models: J48 and RandomForest. The results from WEKA

showed that RandomForest was more accurate, with 68.70% accuracy compared to 66.70% for J48. This indicates that ensemble methods, such as Random Forest, are more effective for complex prediction tasks.

I also used K-Means clustering, which revealed two distinct market segments within the listings. One segment was cheaper, private rooms in Brooklyn, and the other was more expensive, entire homes in Manhattan. These findings provide practical insights that can help Airbnb hosts and the platform itself make better, data-driven decisions about pricing and marketing.

For future work, one could try more advanced models available in WEKA, like neural networks, to see if the accuracy can be improved further. It would also be interesting to include more attributes, like the text from reviews (using text mining), to understand how guest sentiments affect pricing and popularity. Finally, trying different numbers of clusters or other clustering algorithms could reveal even more detailed segments in the market.

9. References

- [1] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
- [2] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.
- [3] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [4] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–15, 2015.
- [5] X. Wu et al., “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2010.
- [6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
- [7] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

- [8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2016.
- [9] A. Ng, “Machine learning and AI via clustering techniques,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 341–356, 2016.
- [10] R. Xu and D. Wunsch, “Clustering algorithms in machine learning,” *IEEE Transactions on Neural Networks*, vol. 25, no. 3, pp. 645–678, 2014.
- [11] D. Dua and C. Graff, “UCI Machine Learning Repository,” UC Irvine, 2019. [Online]. Available: <https://archive.ics.uci.edu>
- [12] F. Li and C. Chen, “Price segmentation and prediction using machine learning: A case study on Airbnb,” *Journal of Artificial Intelligence Research*, vol. 72, pp. 215–230, 2021.
- [13] D. Wang, D. L. Nicolai, and C. R. Zengras, “The spatial structure of Airbnb in New York City,” *Regional Studies, Regional Science*, vol. 8, no. 1, pp. 1–19, 2021.
- [14] A. Silva, H. Moura, and M. Resende, “A machine learning approach for predicting Airbnb rental prices,” in *Proc. IEEE ICDMW*, 2020, pp. 217–224.
- [15] S. Raschka, *Python Machine Learning*, 3rd ed. Packt, 2020.
- [16] S. Tan, *Introduction to Data Mining*, 2nd ed. Pearson, 2018.
- [17] A. Stevens, “Classification techniques in WEKA,” *International Journal of Data Analysis*, vol. 9, no. 2, pp. 55–70, 2019.
- [18] K. Chen and Y. Zhang, “Urban accommodation clustering based on K-Means,” *IEEE Access*, vol. 9, pp. 11023–11035, 2021.
- [19] H. Zhao et al., “Room type prediction using supervised learning,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 4, pp. 610–618, 2021.
- [20] R. Schultz, “Tourism analytics with machine learning,” *Elsevier Tourism Intelligence Journal*, vol. 4, pp. 85–102, 2022.
- [21] Airbnb, “Inside Airbnb: Data on listings,” 2023. [Online]. Available: <http://insideairbnb.com>
- [22] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.