Machine Learning – Assignment 0

Date: 11-10-2025

Group Members:

- Mitzalo Gaston Reyes Polk-Reyes
- Jorge Manuel Velarde Da Silva
- Vishwas Vikram Bheda

---

1. Dataset choice

For this assignment, we have chosen two datasets with varying characteristics.

- Melbourn AirBnB: https://www.openml.org/search?type=data&id=46881
- Wine Reviews: https://www.openml.org/search?type=data&id=46653

Among the differences between them is that one is a wide dataframe, consisting of over a hundred columns of mixed data type, and the other is a much more narrow dataset consisting of only six columns, which nontheless has a free text `description` column which might allow us to obtain further features through the use of text processing. Due to the variety of data contained between each dataset, we expect to be able to try out multiple techniques for data treatment.

2. Dataset summary

AirBnB prices

Property	Description
Rows	18,316
Columns	103
Feature Columns	99
Target Variable	`price_label` (Ordinal)
Attribute Types	Interval, Nominal

Wine Reviews

Property	Description
Rows	84,123
Columns	6
Feature Columns	5
Target Variable	`variety` (Nominal)
Attribute Types	Interval, Nominal

3. Dataset attributes

As one of the datasets has a very large amount of features, for this analysis we have limited it to a smaller set of features.

AirBnB prices

Attribute	Type	Example Values	Range / Categories	Notes
`price_label`	Ordinal	1	[0,9]	Represents an encoding of an Interval type column which is also included in the dataset
`latitude`	Interval	-37.83	[-38.22,-37.48]	
`longitude`	Interval	-37.83	[144.48,145.83]	
`accommodates`	Interval	2	[1,16]	
`bedrooms`	Interval	1	[0,16]	

Wine Reviews

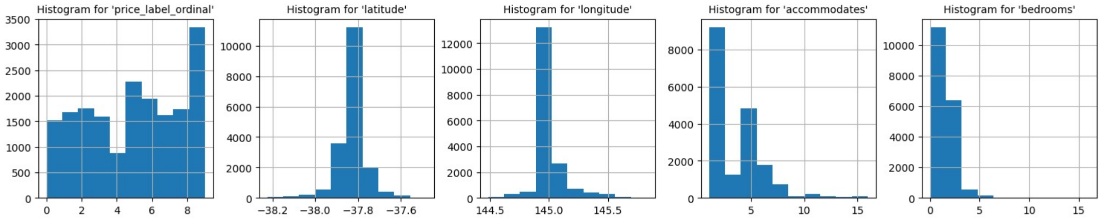
Attribute	Type	Example Values	Range / Categories	Notes
`variety`	Nominal	Cabernet Franc	PENDING	
`country`	Nominal	PENDING	PENDING	
`description`	Nominal	PENDING	PENDING	
`points`	Interval	PENDING	PENDING	
`price`	Interval	PENDING	PENDING	
`province`	Nominal	PENDING	PENDING	

It is important to note that in the **AirBnB** dataset, there are additional price related columns on which the categorical label was based, and so for all downstream tasks, these columns must not be used as they would introduce data leakage into the model.

4. Distributions

We present the histogram of the target as well as a few of the significant features we think might be useful for predicting the target variable.

AirBnB



Wine Reviews

PENDING

5. Data Processing

We have detected multiple columns which will be needed to be treated in different ways, but have not yet determined the best treatment, and which we will experiment during the development of the following assignments as it is currently too early to tell the best technique.

Among these issues, we list a series of possible treatments:

- **Missing values**: Creation of a flag column, imputation (mean, mode, model based, etc), removal of rows, among other methods.
- **Outliers**: Analysis of possible cause (measurement error, possible but rare behavior, etc), removal of rows.
- **Categorical encoding**: One hot encoding, frequency encoding, hash encoding, target encoding, among other methods.