

# Machine Learning Exercise 0

Palle Morris e1160346  
Elmenreich Jan e01526208  
Holzberger Fabian e11921655

March 24, 2020

## 1 Introduction

Two datasets are analyzed, one for classification and a second one for regression. The datasets were chosen such that they have different characteristics. The characteristics and more information about the datasets is listed in the table [1].

Characteristic	Mammographic Mass	House sales
Data Type	Multivariate	Multivariate
Attribute Type	Integer	Integer, String, Real
Associated Tasks	Classification	Regression
Number of instances	961	21436
Number of Attributes	6	20
Missing Values	Yes	No

Table 1: Characteristics of the datasets of choice

We call the mammographic mass dataset 1 and the house sales dataset, dataset 2. Dataset 1 is used for classification and dataset 2 is used for regression. They both are multivariate wherein the difference can be seen in their dimensionality, which is the number of attributes. Here dataset 2 has about 3 times higher dimensionality than dataset 1. Moreover, dataset 2 has a wider variety of attribute types with integer, strings and real numbers where dataset 1 contains merely integers. Additionally, dataset 2 has about 20 times more instances than dataset 1 and doesn't contain any missing values which dataset 1 does.

## 2 Mammographic Dataset

This dataset includes 6 Attributes, we summarize and explain them below

- **BI-RADS Integer** (non-predictive)  
BI-RADS assessment ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy). Can be an indication of how well a CAD system performs compared to the radiologists.
- **Serverity Bool** (target)  
classification by 1 for benign or 0 for malignant
- **Age Integer**  
Age of the specimen
- **Shape Integer**  
mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)

- **Margin Integer**

mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5

- **Density Integer**

mass density: high=1 iso=2 low=3 fat-containing=4

The missing values per attribute can be found table[2]:

BI-RADS	Age	Shape	Margin	Density	Severity
2	5	31	48	76	0

Table 2: Missing values for mammograph dataset

While inspecting the dataset, one instance with a BI-RADS value of 55 was found which, is an error since the range of this attribute is 5 at maximum. We correct this to the value of 5.

The distributions of the dataset attributes can be found in figure [1].

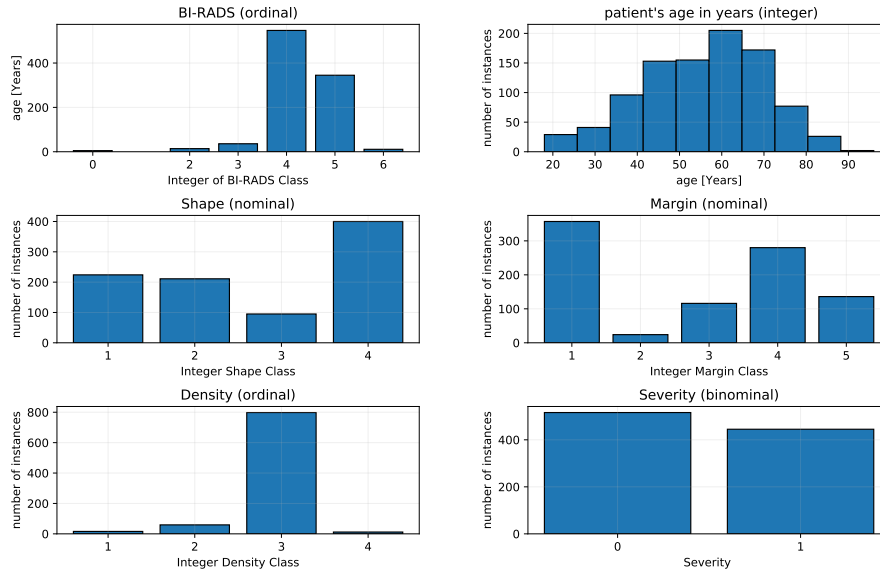


Figure 1: Histogramms of mammographic dataset attributes

In the figure we see that the target attribute severity is binomial. The severe cases are 445 where the non-severe cases are 516 which is slightly not equal. In the BI-RADS attribute that contains ordinal values most values are in class 4 with about 500 followed by class 5 with about 350 values. It can be seen that the age (integer) of the patients is approximately normal distributed around the age 60 with a range of 20 to 90 years. The age is represented as integer value. For the shape attribute (ordinal) class 4 occurs is dominant with about 400 instances followed by class 1 and 2 with slightly more than 200 instances each and class 3 with 100 instances. Next in the margin (ordinal) attribute we see a range of 5 represented classes. Here class 1 and 4 are dominant with about 400 instances for the former and about 300 for the latter one. Lastly in the density (ordinal) attribute only class 3 shows dominance with about 800 occurrences where the other classes 1,2 and 3 occur in less than 100 instances.

### 3 House sales Dataset

This dataset contains 20 attributes + `id`, which we will not count as an attribute. Our target attribute will be, as so often the `price`

- **Price** Real (target)  
Price of each home sold
- **Date** String  
Date of the home sale
- **Bedrooms** Integer  
Number of bedrooms
- **Bathrooms** Real  
Number of Bathrooms
- **Sqft\_living** Integer  
Square footage of the apartments interior living space
- **Sqft\_lot** Integer  
Square footage of the land space
- **Floors** Real  
Number of floors
- **Waterfront** Integer  
A dummy variable for whether the apartment was overlooking the waterfront or not
- **View** Integer  
An index from 0 to 4 of how good the view of the property was
- **Condition** Integer  
An index from 1 to 5 on the condition of the apartment
- **Grade** Integer  
An index from 1 to 13, where 1–3 falls short of the building construction and design, 7 has an average level of construction and design, and 11–13 have a high quality level of construction and design
- **Sqft\_Above** Integer  
The square footage of the interior housing space that is above ground level
- **Sqft\_basement** Integer  
The square footage of the interior housing space that is below ground level
- **Yr\_built** Integer  
The year the house was initially built
- **Yr\_renovated** Integer  
The year of the house's last renovation
- **Zipcode** Integer  
What zipcode area the house is in
- **Lat** Real  
Latitude
- **Long** Real  
Longitude
- **Sqft\_living15** Integer  
The square footage of interior housing living space for the nearest 15 neighbors
- **Sqft\_lot15** Integer  
The square footage of the land lots of the nearest 15 neighbors

In figure [2] we can see the distribution of the target and other attributes.

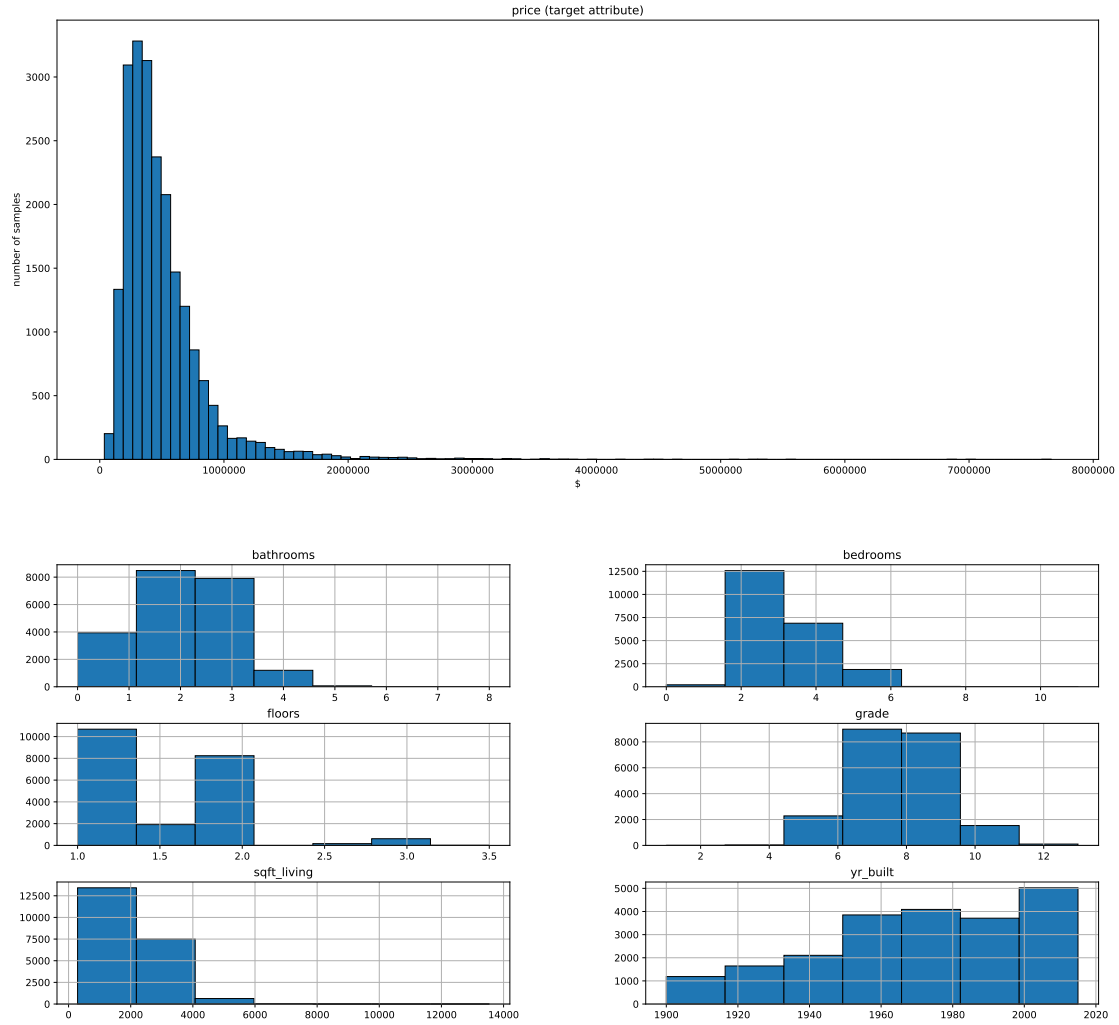


Figure 2: Histogramms of attributes and target attribute

For one instance we encountered a rather high value of 33 bedrooms. After comparing more attributes to different instances, we decided that it could have been a typographical error and changed the value to 3 bedrooms.