# Machine Learning - Assignment 0

Adam Domoslawski (e12241331), Dmytro Kondrashov (e01641210),
Michael Wagner (e00926215)

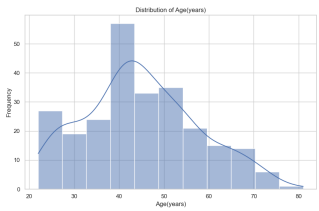
October 17, 2024

We selected two datasets that both tackle important real-world problems, but are significantly different in structure from each other.

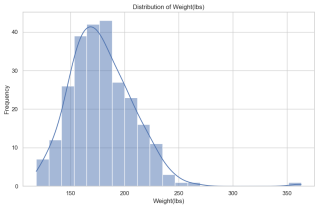
## 1 Regression Dataset: BodyFat

**Motivation:** For regression we selected the *BodyFat* dataset[1]. Obesity related illnesses make up a large proportion medical costs in many developed countries [1]. While studies often classify people into underweight, normal and overweight/obese based on body mass index, a persons percentage of body fat is a much better suited metric for this problem. However, often it is challenging to obtain exact body fat measurements. Various formulas have been proposed to approximate body fat, e.g. by Siri [2] in 1956, however they often relatively unreliable. Hence, machine learning can be an effective tool to obtain a good estimations of a persons body fat.

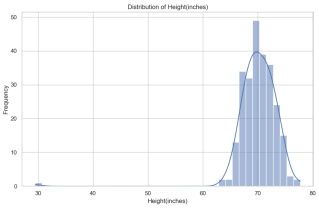
**Description:** This dataset contains 252 samples, each of which containing 14 easily obtainable metrics/attributes such as age, weight, height, density and circumference of various body parts and the associated body fat percentage as a target. None of the samples contains any missing values and the type of all attributes is ratio-scaled. For example, body density is measured in $gm/cm^3$, and we can find a ratio between 1 and 2 because of the square-cube law. Another example is neck circumference, which has a meaningful ratio: 70 cm is twice as large as 35 cm.



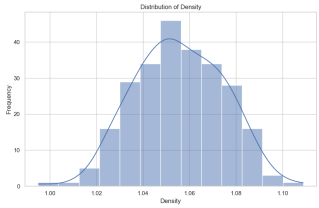

(a) Age distribution



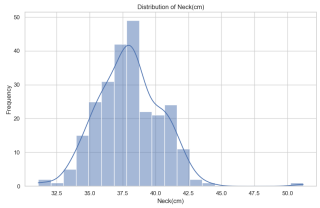

(b) Weight distribution (lbs)



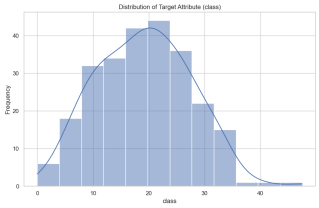

(c) Height distribution (inches)




(d) Density distribution




(e) Neck circumference distribution (cm)




(f) Target group distribution

Figure 1: Distributions of selected attributes and the target of the BodyFat model.

The distribution of some of the attributes is visualized in Figure 1. As we can see, the body fat distribution ranges from 0% to 46%. This must be an incorrect entry since it is impossible for a human to have 0% body fat. Olympic athletes typically have around 15% (male) and 25% (female) body fat on average [3], while the lowest value ever was recorded for a marathon athlete, who had a body fat percentage of 6.4% (±1.3%).

[1] `https://www.openml.org/search?type=data&sort=runs&status=active&qualities.NumberOfClasses=lte_1&id=560`

Another interesting point is the height distribution. One person in the dataset is recorded as being 30 inches (76 cm) tall. This could either represent a person of short stature or be an error in the data. It is better to check and potentially remove these samples during the pre-processing step to remove incorrect samples, reduce outliers and ensure a more uniform data distribution for modeling.

## 2 Classification Dataset: Porto-seguero

**Motivation:** For (binary) classification we selected the *Porto-seguero* dataset [2]. Car insurance can be very expensive, however insurance companies need to select a reasonable price in order to make a profit. One way to make insurances more affordable while encouraging safe driving is to predict the likelyhood for drivers to get into an accident and provide *safe* drivers with better offers. However, prediction the likelyhood of accidents is not an easy task, which is why Porto Seguro hosted a competition in 2017 to find the best approach to predict the probability of drivers filing an insurance claim within the next year [4]. The training set from that competition can be used for binary classification.

**Desrciption** This dataset differs a lot from the BodyFat dataset, both in terms of size and in terms of structure. First, it is much larger, containing more than half a million samples, 37 different attributes and a boolean flag specifying whether the specific driver ended up filing an insurance claim. Second, in contrast to the BodyFat dataset, this dataset contains attributes of various different types of data: We identified 25 nominal, (12 of which were binary), 5 ordinal and 7 interval attributes.

**Missing Values** Further, there exist multiple attributes where some samples do not contain a value (many samples with some missing values). To handle the missing values we took two different approaches. For nominal attributes, where only a relatively small number of samples was missing a value, we simply removed the sample from the dataset. However, for two attributes there was a proportionally large number of samples with missing values. Removing all these samples would drastically reduce the size of the dataset, so we instead replaced those missing values with the mean over all existing values for these attributes, which was possible since both parameters are ordinal.

We visualize the distribution of all ordinal and (most) interval variables as well as the distribution of the binary target in Figures 2, 3 and 4. Visualizations for the remaining 26 attributes are omitted for space reasons. Notably, most of the ordinal and interval attributes either have a relatively even distribution or one that looks similar to a Gaussian distribution, with the exception of attribute ps_ind_14, where most samples take the value 0. Finally, there are significantly more samples where the target value is false, which is to be expected, considering people generally tend to not get into accidents every year.

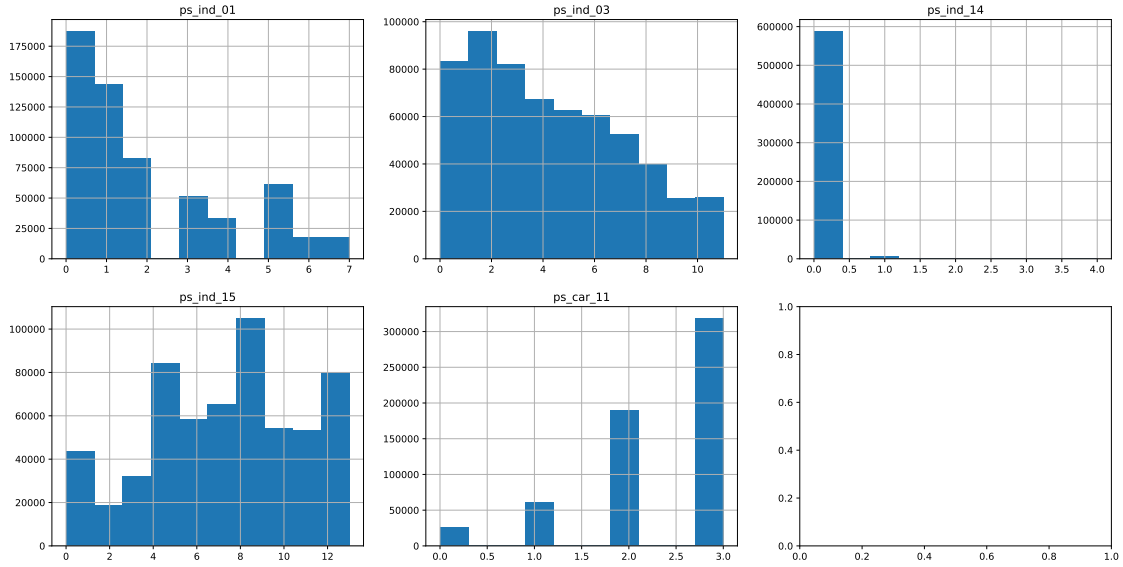



Figure 2: Distributions of ordinal attributes

[2] https://www.openml.org/search?type=data&sort=version&status=any&order=asc&exact_name=porto-seguro&id=42206

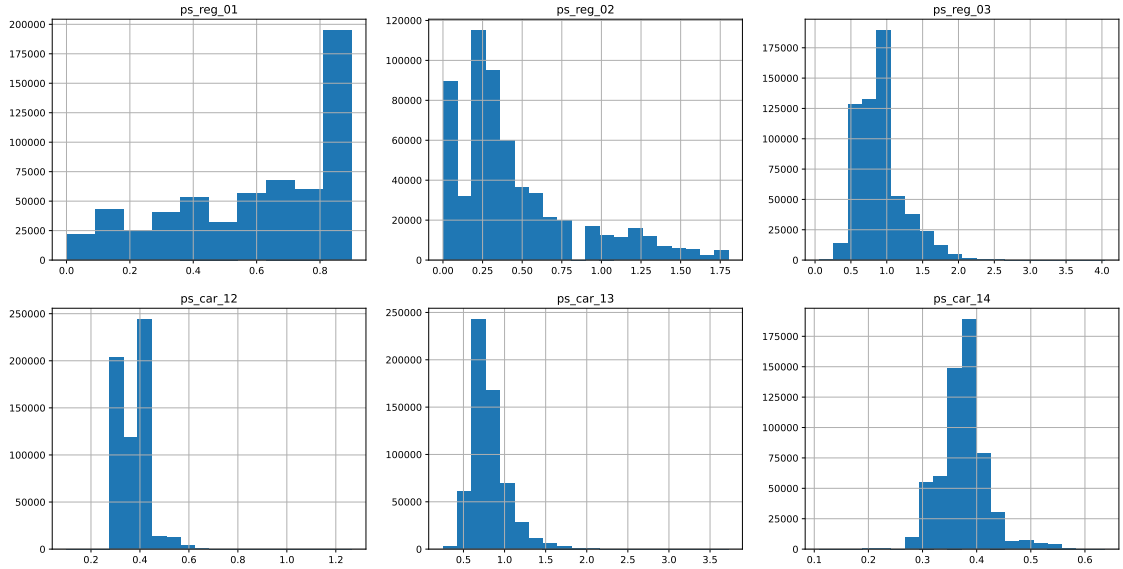



Figure 3: Distributions of interval attributes

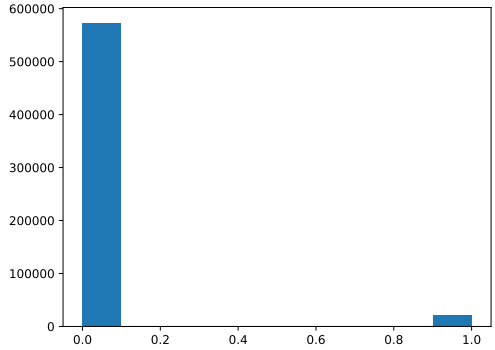


Figure 4: Distribution of target value

# References

[1] G. Colditz, "Economic costs of obesity," *The American Journal of Clinical Nutrition*, vol. 55, no. 2, pp. 503S–507S, 1992.

[2] W. E. Siri, "In advances in biological and medical physics," *London and New York, Academic press Inc.*, vol. 4, pp. 239–280, 1956.

[3] S. J. Fleck, "Body composition of elite american athletes," *The American journal of sports medicine*, vol. 11, no. 6, pp. 398–403, 1983.

[4] i. Addison Howard, AdrianoMoala, "Porto seguro's safe driver prediction," 2017. [Online]. Available: https://kaggle.com/competitions/porto-seguro-safe-driver-prediction