Assignment 0

Jonas Kompauer, 11776872
Lukasz Sobocinski, 12123563
Florian Lackner, 11704916

Dataset 1: Medical Cost Personal Datasets (Regression)

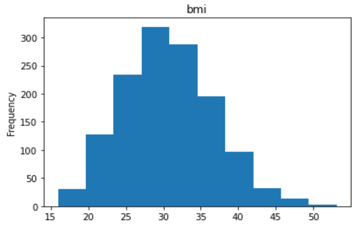
https://www.kaggle.com/mirichoi0218/insurance/version/1

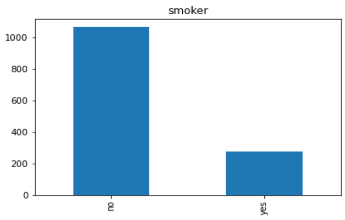
We chose this data set because it is an interesting topic and fits well for this Lecture as it has, in contrast to our second dataset, only 6 input attributes, and the sample size is rather small, but not too small, so we have enough data for learning.

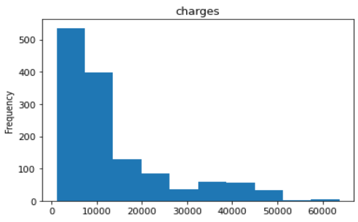
Characteristics of data set

The dataset contains 1338 samples with each having 7 attributes, where one attribute is the target attribute.

The attributes have different data Types: three nominal attributes (sex, smoker, region), three ratio attributes (age, children and charges) and one interval attribute (bmi) as the ratio of bmi values does not make sense.







Here we see the Histograms of a few attributes, the bmi, smoker and charges attributes. “Bmi” is, more or less, normal distributed while “charges” is more hyperbolic, as there are many lower values and only some higher values. “Smoker” has 5 times more “no” values as “yes” values.

Target attribute

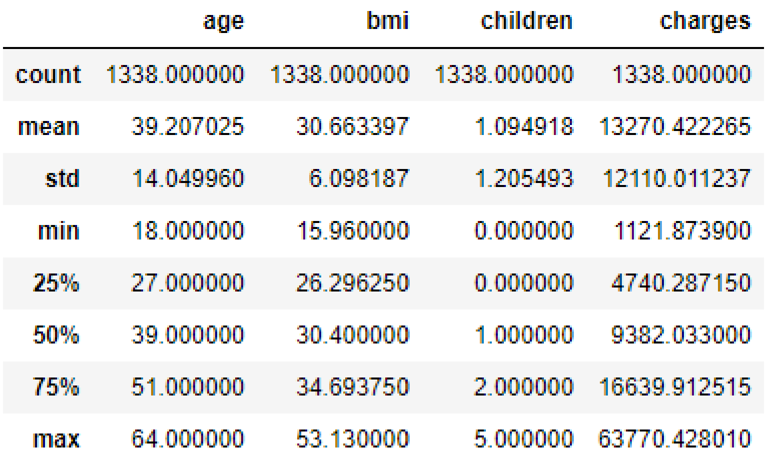
The target attribute is “charges” is a ration data type and the values range from 0 to 16884.924, this is important to have an idea on what output to expect of an algorithm. If our algorithm returns e.g. negative values, too many large values, we should further investigate it because there could be errors in the algorithm. Also if the distribution of the target attributes from a Test set is not similar, it is another hint that the algorithm could not be working properly.

Categorical data

We only have nominal categorial data attributes (sex, smoker, region) as the values of these attributes can’t be compared. “Sex” and “smoker” both only have two possible values, “female”/”male” and “yes”/”no”. Region has four different possible values: “southwest”, “southeast”, “northeast” and “northwest”. This is important to know as we need to pre-process these categories e.g. with a “One hot encoding”, before being able to work with these input values. We can only do frequency distribution, as nominal data types cannot be compared.

Numeric values

The ranges of the interval attribute (bmi) and the ratio attributes (age, children, charges) are shown in the table below



	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Other important aspects

We probably need to standardise the values, depending on the algorithm which we will use, to ensure that the different scales do not impact the algorithm

Dataset 2: Speed Dating (Classification)

https://www.openml.org/d/40536

We chose this data set because of the many attributes and therefore exact prediction of the matches. Furthermore there is data from both perspectives (both persons within the date). Moreover it shows human behaviour and points out stereotypes at dates.

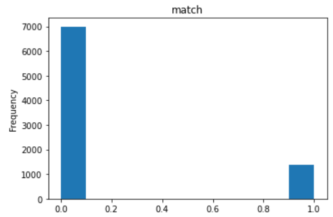
Characteristics of data set

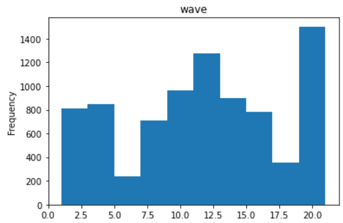
8378 samples with each having 121 attributes. 1.81% of the values are missing in total. 7330 rows have at least one value missing. There is an attribute has_null which shows if a row contains null values.

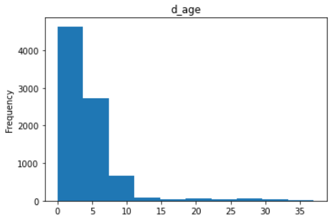
The study was made in several waves. Each wave is a group of people coming to the Speed Dating event and making short talks to determine if they like each other. If both persons had a positive opinion about the other one, then it is marked in the “match” variable as “1”. The study was conducted considering only male and female relationships. In the dataset we can observe some groups of similar data:

1. Basic data about the person, like age or gender.
2. Information how participants evaluate their character and what partner's character they prefer. Moreover, we know what those values are of the person with whom they talked.
3. Information about their hobbies and the other person's hobbies.
4. Ratings how participants estimated their chances of matching with someone.

More information about available attributes can be found in the Appendix.







Target attribute

The target attribute is "match" which has nominal data types with possible values only being "0" meaning "no match" or "1" meaning "match". In this dataset a target value "0" appears five times more than a "1".
This is important to know, because when learning, our algorithm should only output one of these two values, and we should expect a similar distribution when working with a test set.

Categorical data

In this dataset exist nominal and ordinal data types, but there are far more attributes with ordinal data types. The list of all data types is in the appendix. This is important to know as we need to pre-process these categories e.g. with a "One hot encoding", before being able to work with these input values.

Numeric values

The attributes with numeric data types all have different kinds of ranges so it would be important to standardise these values to be able to work with them.

Other important aspects

For preprocessing we may need to change data from the "field" attribute, as there are entries with "law" and also "Law" which should be in the same category.

Appendix

Dataset 1 Features:

Nominal: sex, smoker, region
Ordinal:
Interval: bmi
Ratio: age, children, charges

Dataset 2 Features:

Nominal: gender, race, race_o, field, samerace, has_null, field, decision, decision_o, match

Ordinal: d_d_age, d_importance_same_race, d_importance_same_religion, d_pref_o_attractive, d_pref_o_sinsere, d_pref_o_intelligence, d_pref_o_funny, d_pref_o_ambitious, d_pref_o_shared_interests, d_attractive_o, d_sincere_o, d_inteligence_o, d_funny_o, d_ambitious_o, d_shared_o, d_attractive_important, d_sincere_important, d_attractive_partner, d_sincere_partner, d_intelligence_partner, d_funny_partner, d_shared_intersts_partner d_inteligence_important, d_funny_important, d_shared_interest_important, d_attractive, d_sincere, d_inteligence, d_funny, d_shared_intersts_partner, d_sport, d_tvsports, d_exercise, d_dining, d_museums, d_art, d_hiking, d_gaming, d_clubbing, d_reading, d_tv, d_theater, d_movie, d_concerts, d_music, d_shopping, d_yoga, d_interests_correlate, d_expected_happy_with_sd_people, d_expected_num_interested_in_me, d_expected_num_matches, d_like, d_guess_prob_liked

Interval: importance_same_race, importance_same_religion, pref_o_attractive, pref_o_sinsere, pref_o_intelligence, pref_o_funny, pref_o_ambitious, pref_o_shared_interests, attractive_o, sincere_o, intelegence_o, funny_o, ambitious_o, shared_o, attractive_important, sincere_important, inteligence_important, funny_important, shared_interest_important, attractive, sincere, intelligence, funny, attractive_partner, sincere_partner, inteligence_partner, funny_partner, sport, tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv, theater, movie, concerts, music, shopping, yoga, interests_correlate, expected_happy_with_sd_people, expected_num_interested_in_me, expected_num_matches, like, guess_prob_liked,

Ratio: age, age_o, d_age, met