Machine Learning

TU Wien

Exercise 0: Dataset description

Group 6
Diogo Braga (E12007525)
Enrico Coluccia (E12005483)
Matthias Kiss (01218325)

Short introduction of our team members:

Diogo: My name is Diogo Braga, I am Portuguese and I am currently developing my master's thesis, with Machine Learning being the main research area.

Enrico: My name is Enrico Coluccia I am currently in my final year of MSc at Unisalento (Lecce, Italy) where I am studying Computer Engineering.

Matthias: My name is Matthias Kiss, and I am currently doing my PhD at the institute of Chemical, Environmental & Bioscience Engineering at TU Wien, where I work on computational fluid dynamics simulations.

Datasets

- Breast Cancer Data Set (Classification)
- Metro Interstate Traffic Volume Data Set (Regression)

Datasets were chosen with care to present different characteristics, so that during the development of the work we deal with different problems. The classification data set has a lower number of instances (286), 9 attributes and two targets (binary). It also has missing values, which is common in real life data, so it is a good choice to include this condition. On the other side, the regression data set has a higher number of instances (48204), also 9 attributes and a continuous numeric target. The missing values show as *N/A*, which is challenging to check whether there are missing values or not. It is also going to be interesting to compare the influence of the same number of attributes in data sets with so much difference in number of instances. Both datasets are included in interesting areas such as medicine and transportation organization, which helped to choose them.

Breast Cancer Data Set

We chose this set as our classification data set. There are two classes (binary classification): “no-recurrence-events” and “recurrence-events”, that describe whether the patient’s cancer reappeared after treatment. The other 9 attributes contain general information about the patients themselves as well as more specific information about their individual cancer diagnoses. Using this information the goal is to classify whether a patient will have breast cancer again, or not.

Preprocessing: The missing values in the dataset were marked with ‘?’. To check them, we preprocess the data by converting the missing data into NaN.

Data set characteristics:

1) Description of attributes:

Attributes	Values	Data type
Class	no-recurrence-event, recurrence-event	nominal quantity
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99	interval quantity
menopause	lt40, ge40, premeno	nominal quantity
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59	interval quantity
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39	interval quantity
node-caps	yes, no	nominal quantity
deg-malig	1, 2, 3	ordinal quantity
breast	left, right	nominal quantity
breast-quad	left-up, left-low, right-up, right-low, central	nominal quantity
irradiate	yes, no	nominal quantity

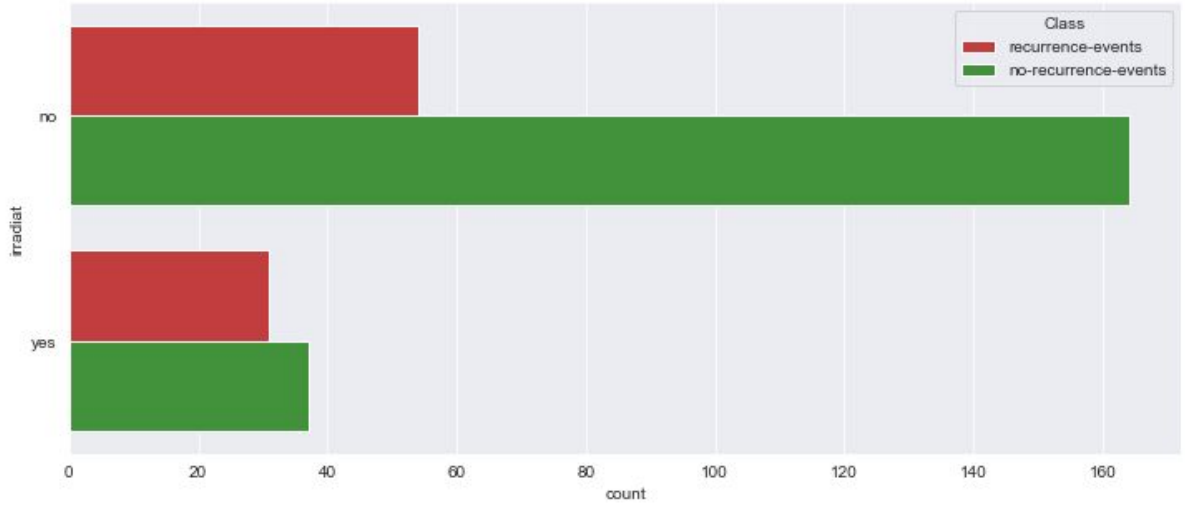
- **Class:** Describes if a patient had recurrent tumors. Nominal quantity, because there is no measurable relation between the two classes.
- **Age:** Age is listed in intervals of 10 years. We would argue it is not a ratio quantity, but an interval quantity, because despite age having a zero-point, no meaningful ratios can be built between the intervals. A person that is in the 20-29 year range is not twice as old as one in the 10-19 year range but somewhere between 1 day and 20 years older. Therefore not all mathematical operations are allowed.

- **Menopause:** This is another nominal quantity. The three possible values can not be put into mathematical relationships.
- **Tumor-size:** Greatest diameter in (mm) of a tumor. Interval quantity, same argumentation as with the age attribute.
- **Inv-nodes:** Number of lymph-nodes in close proximity of the tumor. Interval quantity, same argumentation as with the age attribute.
- **Node-caps:** Indicates whether there are metastases in the surrounding lymph nodes. This is a nominal quantity.
- **Deg-malign:** Describes how bad the cancer is. This is an ordinal quantity, because the values can be ordered (1-best to 3-worst), however there is no distance defined between the values.
- **Breast:** Tumor in left or right breast, nominal quantity.
- **Breast-quad:** Location of tumor in breast, nominal quantity.
- **Irradiate:** If the patient underwent radiation therapy, nominal quantity.

2) Distribution of values

There are a total of 286 samples per attribute in this dataset. Of the 9 attributes and 1 class there are 6 nominal quantities, 1 ordinal quantity, 3 interval quantities and no ratio quantities. A total of 9 values are missing, 8 are in the attribute “Node-caps” and one in breast-quad.

3) Histograms of attributes:



We think the relationship between “no-recurrence” and radiation therapy is very interesting. Intuitively we would assume that radiation therapy significantly increases the chance of having non-recurring cancer. However it seems like the chances are procentually better without that therapy. We will investigate if reasons for this statistic can be found in the data. For example if radiation therapy is only used in people with larger tumors or a higher malignity to begin with.

Metro Interstate Traffic Volume Data Set

This data set contains information related to hourly interstate 94 Westbound traffic volume for Minnesota DoT ATR station 301. The main goal is, with access to the data, to establish a regression that fits the data and is able to predict the traffic volume for this specific station. This is considered a regression problem because we are dealing with a continuous variable target, the traffic volume.

There was no preprocessing necessary for this step of the analysis of the data set.

Data set characteristics:

1) Description of attributes:

Attributes	Values	Data type
Traffic Volume (Regression Variable)	Integer ∈[0, 7280]	ratio quantity
Holiday	None, Columbus Day, Veterans Day, Thanksgiving Day, Christmas Day, New Years Day, Washington's Birthday, Memorial Day, Independence Day, State Fair, Labor Day, Martin Luther King Jr Day	nominal quantity
temp	Float ∈[0, 310.07]	ratio quantity
rain_1h	Float ∈[0, 9831.3]	ratio quantity
snow_1h	Float ∈[0, 0.51]	ratio quantity
clouds_all	Integer ∈[0, 100]	ratio quantity
weather_main	Clouds, Clear, Rain, Drizzle, Mist, Haze, Fog, Thunderstorm, Snow, Squall, Smoke	nominal quantity
weather_description (alternative)	38 unique strings describing the weather in more detail	nominal quantity
date_time	yyyy-mm-dd hh:mm:ss between 2012-10-02 09:00:00 and 2018-09-30 23:00:00	interval quantity

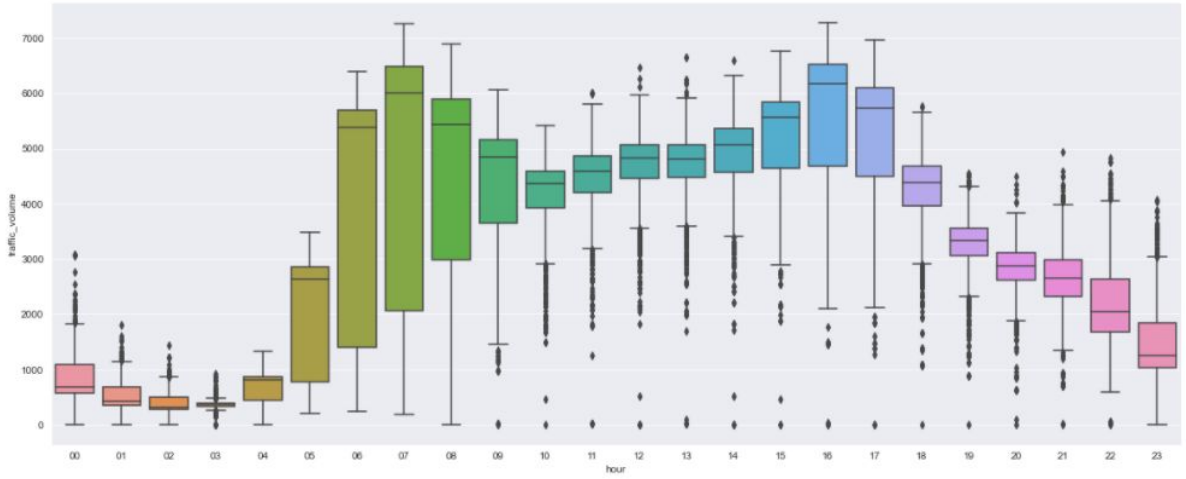
- **traffic_volume:** Hourly I-94 ATR 301 reported westbound traffic volume. This is the regression variable. It is a ratio quantity.
- **holiday:** US National holidays plus regional holiday, Minnesota State Fair. This is a nominal quantity.
- **temp:** Average temperature over one hour in Kelvin. This is a ratio quantity, as Kelvin has a defined zero-point.
- **rain_1h:** Amount in mm of rain that occurred in the hour. This is also a ratio quantity.

- **snow_1h:** Numeric Amount in mm of snow that occurred in the hour. Also a ratio quantity.
- **clouds_all:** Numeric Percentage of cloud cover. This is a ratio quantity.
- **weather_main:** Short textual description of the current weather, nominal quantity.
- **weather_description:** Longer textual description of the current weather, nominal quantity.
- **date_time:** Date and time of the data collected in local CST time. This is an interval quantity.

2) Distribution of values

There are 48204 samples in this data set. A total of 8 attributes is measures, one of which is the ration quantity traffic_volume that is the regression variable. There are 3 nominal, 0 ordinal, 1 interval and 4 (+1 regression variable) quantities. No missing values were indicated by the use of a special value as was done in the breast cancer data set.

3) Histograms:



In this boxplot it is possible to see the distribution of the traffic volume by the hours of the day. It is easily visible that the volume of traffic is higher between 6 am and 7 pm, which can be explained by the usual working hours of people. It is interesting to see the actual data presented in an easily understandable way, and so, like this chart, we built others to try to analyze the data better and also build better conclusions.