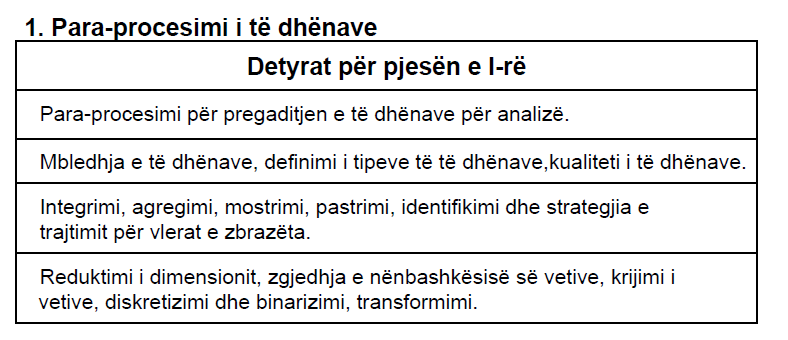
**Kerkesat:**

What is Data? Data is collection of data objects and their attributes. An attribute is a property or characteristic of an object. Attribute is also known as entity, field or feature.

What are attribute values? Attribute values are numbers or symbols assigned to an attribute.

Distinctions between attributes and attribute values: - Same attribute can be maped to different attribute values. Example: height can be measured in feet or meters. - Different attributes can be maped to the same set of values. Example: Attribute values for ID and age are integers, but properties of attribute values can be different. ID has no limit but age has a maximum and mimimum value.

**Attributes types:**

There are 4 different types of attributes:

- Nominal – examples: ID number, eye color, zip codes

- Ordinal – examples: rankings (e.g. of movies on a scale from 1 to 10), grades, height in {tall, medium, short}

- Interval – examples: calendar dates, temperatures in Celcius or Faranheit

- Ratio – examples: temperature in Kelvin, counts, time, length

**Discrete and Continuos Attributes:**

● Discrete Attribute:

Has only a finite set of values. Example: zip codes. Often represented as integer variables. *Note: Binary attributes are a special case of discrete attributes.*

● Continuos Attribute:

Has real numbers as attribute values. Examples: temperature, height or weight. Practically, real values can only be measured and represented using a finite number of digits. Continuos attributes are typically represented as floating-point variables.

**Data Quality:**

Example of data quality problems:

● Noise and outliers

● Missing values

● Duplicate data

● Noise refers to modification of original values. Outliers are data objects with characteristics that are different than most of the other data objects in the data set. ● Missing values – reasons for missing values: information is not collected (e.g., people decline to give their age or weight), attributes may not be applicable to all cases (e.g., annual income is not applicable to children). How to handle missing values? – - delete rows with missing values, - impute missing values with Mean/Median (for continuos variables) - impute missing values with most frequent value (for categorical variables) - predict missing values)

● **Duplicte Data:**

Data set may include data objects that are duplicates. Example: same person with multiple email addresses Data Cleaning – process of dealing with duplicate data issues.

**Data Preprocessing:**

- Aggregation

- Sampling

- Dimensionality Reduction

- Feature subset selection

- Feature creation

- Discretization and Binarization

- Attribute Transformation

● **Aggregation**

Combining two or more attributes (or objects) into a single attribute (or object). Purpose: data reduction, change of scale, more “stable” data

● **Sampling**

It is used because processing the entire set of data is time consuming. Using a sample will work almost as well as using the entire data set, if the sample is representative. A sample is representative if it has approximately the same property as the original set of data.

● **Dimensionality Reduction**

Purpose – reduce amount of time and memory required by data mining algorithms, allow data to be more easily visualized.

● **Feature subset selection**

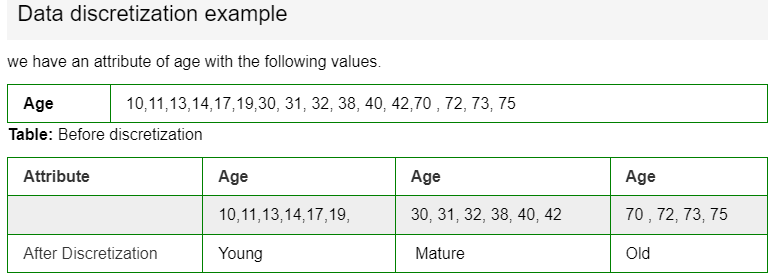
Another way to reduce dimensionality of data - Redundat features: duplicate much of the information contained in one or more attributes, - Irrelevant features – contain no information that is useful.

● **Feature Creation**

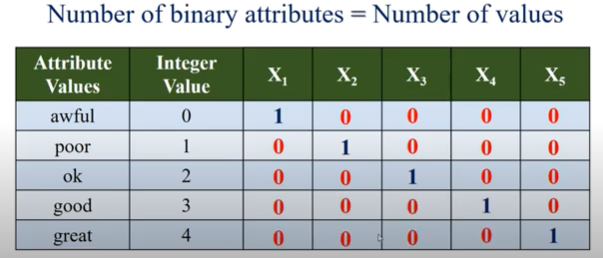
Create new attribute that capture important info in much more efficiently than original attributes.

● **Discretization and Binarization**

Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value.



Binarization is the process of transforming data features of any entity into [vectors](https://deepai.org/machine-learning-glossary-and-terms/vector) of binary numbers. Binarization is the process of dividing data into two groups and assigning one out of two values to all the members of the same group.



● **Attribute Transformation:**

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

**Mbrojtja:**

Per realizmin e projektit kemi perdorur veglen Anaconda dhe Notebook Jupyter, si dhe gjuhe programuese per te dhena Python.

**Mbledhja e te dhenave**: eshte realizuar duke marrur nje dataset te gatshem nga website kaggle.com. Ky dataset ka te beje me te dhenat e motit ne Australi. Kemi ruajtur lokalisht datasetin si .csv file dhe e kemi hapur dhe lexuar ate permes librarise pandas dhe e kemi ruajtur ne nje variabel te quajtur traindata.

**Definimi i tipeve te te dhenave:**

Date object

Location object

MinTemp float64

MaxTemp float64

Rainfall float64

Evaporation float64

Sunshine float64

WindGustDir object

WindGustSpeed float64

WindDir9am object

WindDir3pm object

WindSpeed9am float64

WindSpeed3pm float64

Humidity9am float64

Humidity3pm float64

Pressure9am float64

Pressure3pm float64

Cloud9am float64

Cloud3pm float64

Temp9am float64

Temp3pm float64

RainToday object

RainTomorrow object

Ne datasetin tone kemi tipe te ndryshme te te dhenave si: string, date, double.

**Kualiteti i te dhenave:**

Perfshin: ● Noise and outliers (Noise dhe Outliers perfshihen ne fazen e II) ● Missing values – Identifikimi dhe strategjia e trajtimit per vlerat e zbrazeta -

Pandas DataFrame dropna() function is used to remove rows and columns with Null/NaN values.

● Duplicate Data – data cleaning – traindata.duplicated() i kqyr nese dy rreshta jane te njejte

traindata = traindata.drop\_duplicates() – i largon te gjitha duplikatet

**Integrimi** – e kemi realizu permes importimit te librarise pandas. Fillimisht e kemi ruajtur .csv file lokalisht, pastaj e kemi hapur ate permes funksionit open() dhe permes librarise pandas dhe funksionit read\_csv() e kemi lexuar fajllin .csv dhe e kemi ruajtur ne variablen X.

**Agregimi** – dy kolona i kemi bashkuar (agreguar) ne nje kolone te vetme per te bere reduktimin e dimensionit.

**Mostrimi** – kemi treguar shembuj se si mund te behet mostrimi me ane te funksionit built in sample() i cili gjeneron ne menyre random nje numer te caktuar te rreshtave te datasetit. Mostrimi mund te perfshije edhe nje perqindje te datasetit p.sh. 25% me ane te opsionit frac te funksionit sample(). Ne fund kemi shtuar nje kusht per te pare nese perqindja e mostres se marre eshte me te vertete aq sa kemi definuar ne funksion

**Reduktimi i dimensionit:** perfshin funksionin corr(), shape(), describe(). Reduktimi i dimensionit lidhe me trajtimin e vlerave null.

**Zgjedhja e nenbashkesise se vetive: ??**

**Krijimi i vetive:** me shtu nje kolone te re – temperatura mesatare e dites.

**Diskretizimi dhe binarizimi:** nuk i kemi pare te nevojshme per tu aplikuar ne projektin tone.

**Transformimi:** kemi transformuar kolonen temperatura mesatare prej celcius ne kelvin.