VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

AIML 231/DATA 302— Week 5

Data Preprocessing

Dr Bach Hoai Nguyen

School of Engineering and Computer Science

Victoria University of Wellington

Bach.Nguyen@vuw.ac.nz

Week Overview

- Introduction of Data Preparation
  - What data preparation include
  - Why data preparation
  - Avoid data leakage

- Data Preprocessing
  - Categorical Data Encoding
  - Normalisation
  - Discretisation
  - Impute missing values

- Feature Manipulation
  - Dimensionality Reduction
  - Feature Construction
  - Feature Selection

Data Preprocessing/Preparation

- Prepare the final data set(s) for modelling
- Takes over 80% of time and effort in the project
- Five steps:
  - Data Selection: determine data sets to be used, select features, select instances
  - Data Cleaning: to correct, impute, or remove erroneous values, missing values
  - Data Construction: constructive data preparation operations, e.g. feature construction, instance generation, feature transformation
  - Integrate data: create new records or values by combined from multiple data source, merge data from different sources, aggregations
  - Format data: re-format data, convert to format convenient for modelling

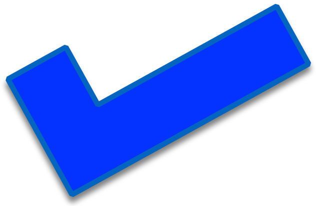
Why Data Preprocessing?

- Data in the real world:
  - incomplete: missing attribute values
  - inconsistent: “03/07/2015”, “March 07, 2015”
  - noisy: containing errors or outliers, gender=“Male”, pregnant = “Yes”
  - large-scale/big data: with a large number of features and instances
  - different types: numeric, nominal, text, Web data, images, audio/video
- Different ML tools use different data formats; Different ML methods have different requirements
- Garbage in, garbage out

Data Leakage

- A Problem with naive data preparation - data leakage
- Information/knowledge about the holdout dataset, e.g. a test dataset, leaks into the data used to train the model
- result in an incorrect estimate of model's perdiction performance

Data Preparation->Data Splitting-> Modelling 

Data Splitting-> Data Preparation-> Modelling 

Data Preprocessing

- Different types of data:
  - Numerical data: discrete (integers) vs continuous
  - Categorical data: nominal (colours) vs ordinal (education level)
  - other/special types of data (multi-media data):Text data, hyperlink data, image data

- Encoding categorical data: convert categorical data to numerical value
- Nomalisation/Scaling: transform columns/rows to a consistent set
- Discretisation: convert a numeric attribute to a nominal attribute
  - e.g. Temperature attribute from {50, 80} to {low, high}
- Impute missing values

Categorical Data Encoding Scheme

- Categorical variables : contain label values rather than numeric values

- One Hot Encoding:
  - Nominal data
  - for each unique value in a categorical column, a new column is added
  - *sklearn.preprocessing.OneHotEncoder*

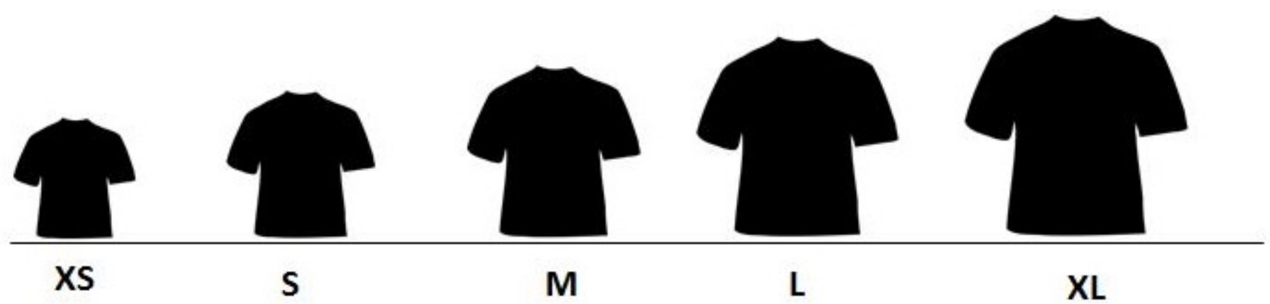
Country
USA
UK
USA
France
USA
UK

dummy variables

USA	UK	France
1	0	0
0	1	0
1	0	0
0	0	1
1	0	0
0	1	0

Encoding Ordinal Variables

- Ordinal data: categorical data that have a natural rank order



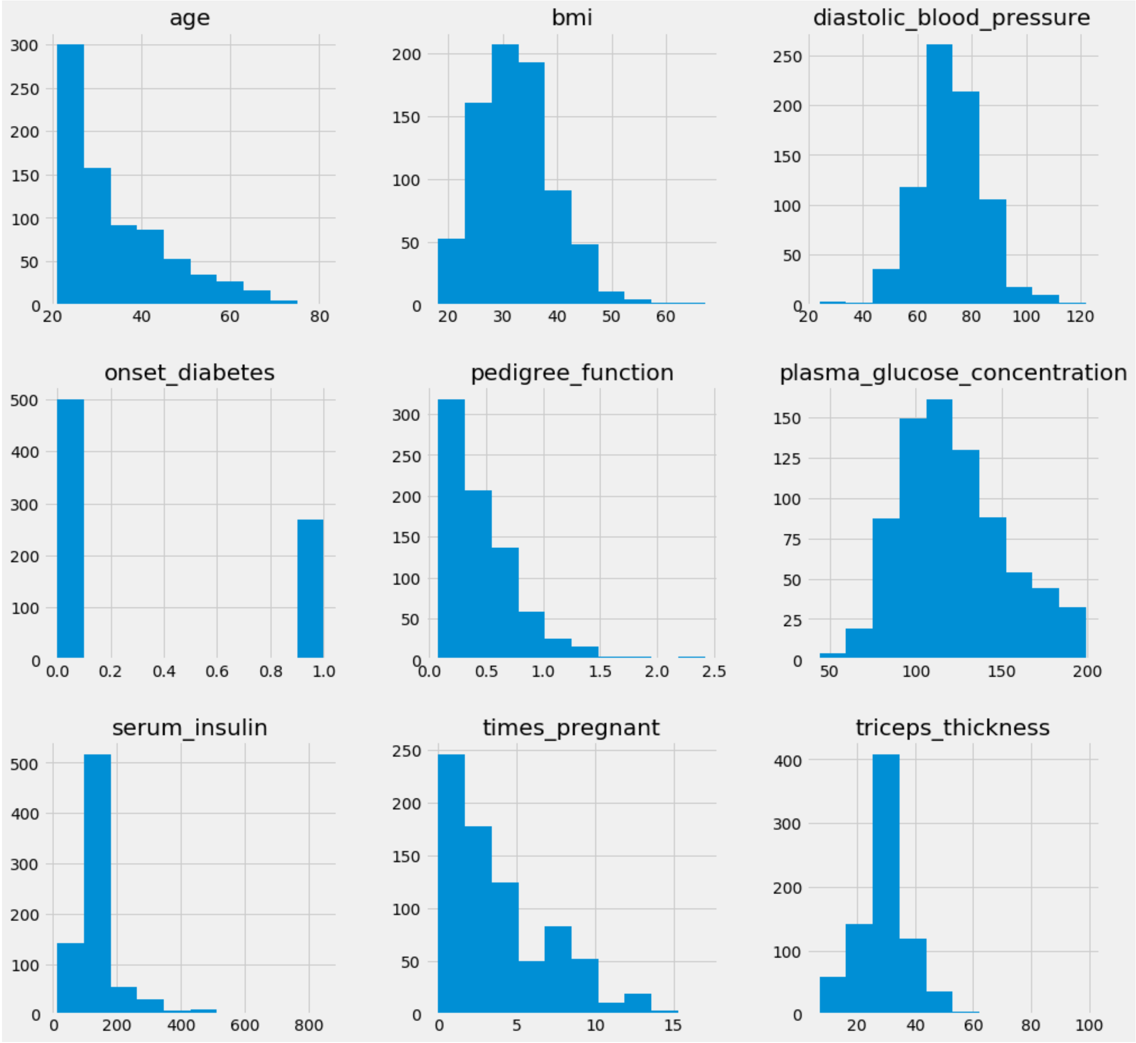
- Ordinal encoding: assign integers to labels in certain order

Original values	Encoding values
XS	0
S	1
M	2
L	3
XL	4

- *sklearn.preprocessing.OrdinalEncoder*

Normalisation/Scaling

- Numerical data: feature values in different ranges
- Some machine learning methods e.g. KNN, SVM, gradient descent, are affected greatly by the scale of the data
- Normalisation transforms columns and/or rows to a consistent set of rules
- a common form - transform all features to be between a consistent and static range of values, e.g.[0, 1]



Example of variables with vastly different scales

Min-Max Normalisation/Scaling

- To the range [0, 1]:
  - the min are all zeros and the max values are all ones

$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

- To a pre-defined range [New_{min}, New_{max}]:

$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}(New_{max} - New_{min}) + New_{min}$$

- Use *sklearn.preprocessing.MinMaxScaler*

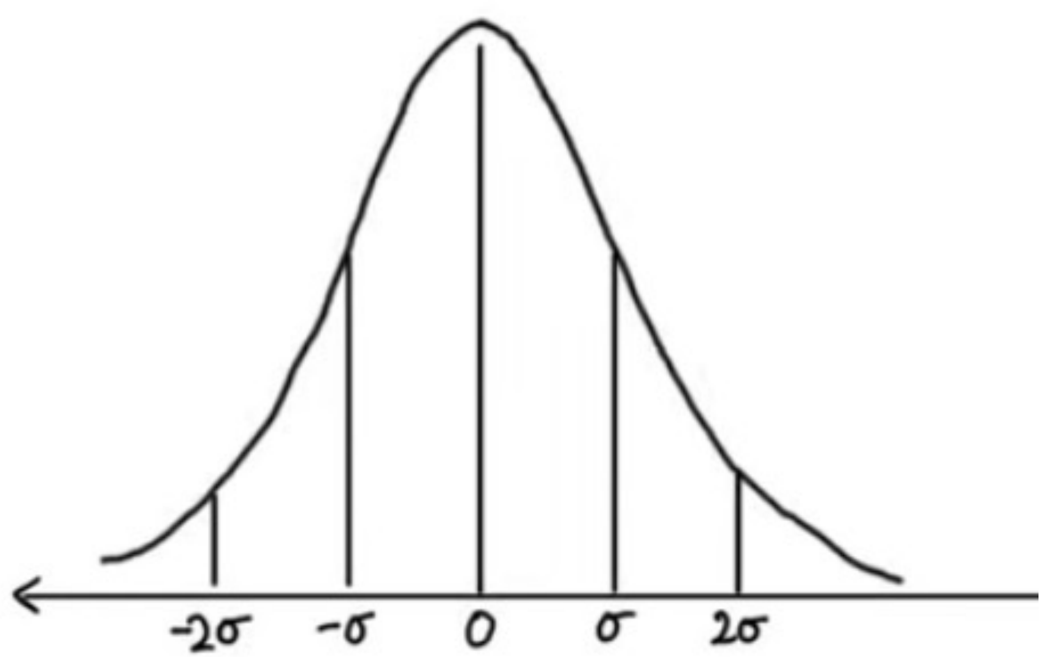
Z-Score Standardisation

- Center scaling
- Center values around a mean of zero and a standard deviation of one utilising the statistical idea - a z-score/standard score

$$z = (x - \mu)/\sigma$$

μ is the mean, σ is the standard deviation of the feature

- Use *sklearn.preprocessing.StandardScaler*



Normalisation or Standardisation?

- Standardisation can give values that are both positive and negative centered around zero
- Normalisation makes different variables to have the same range
- If the distribution is normal, then it should be standardised, otherwise => normalise
- If in doubt => normalise
- might be a good idea to have a mixture of standardised and normalised variables=> standardised followed by normalised

Discretisation

- Discretisation: a process of converting continuous values such as price, age, and weight into discrete intervals
- Some algorithms prefer/require categorical inputs, e.g. DT, rule-based algorithms
- For data smoothing, handle outliers

Two types:

- Unsupervised discretisation - does not depending on class label
  - *sklearn.preprocessing.KBinsDiscretizer*
- Supervised discretisation - depends on class label
  - 1RD, entropy-based

Discretisation: Equal-Width/Uniform

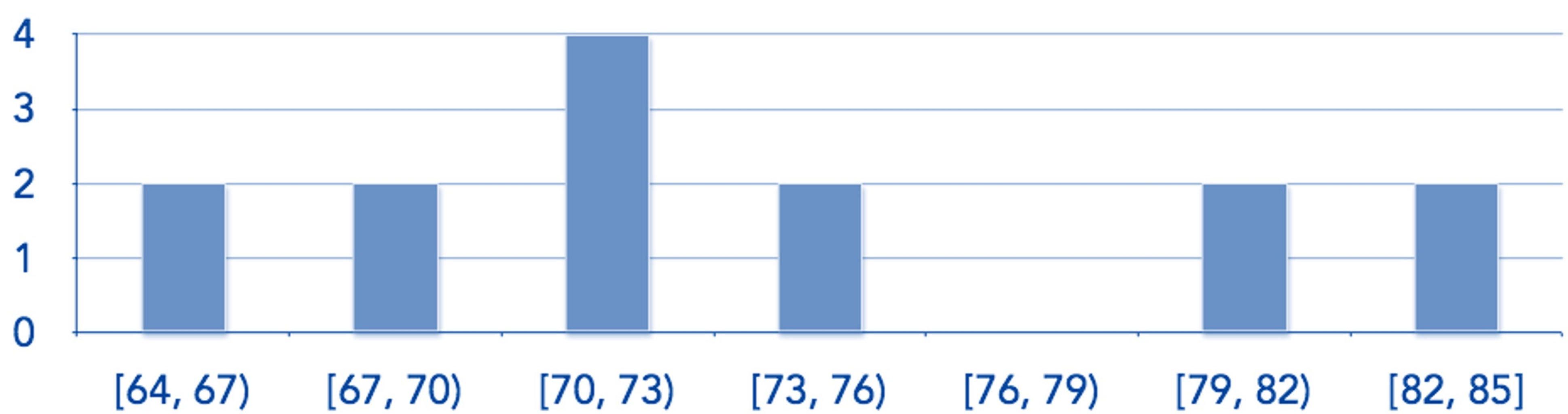
Convert a numerical attribute to an ordinal attribute with N possible values

- Find the Maximum and Minimum values of the attribute
- Divides the range [Min, Max] into N intervals of equal size
- The width of intervals: W=(Max - Min)/N

- KBinsDiscretizer(n_bin=7, encode='ordinal', strategy='uniform')

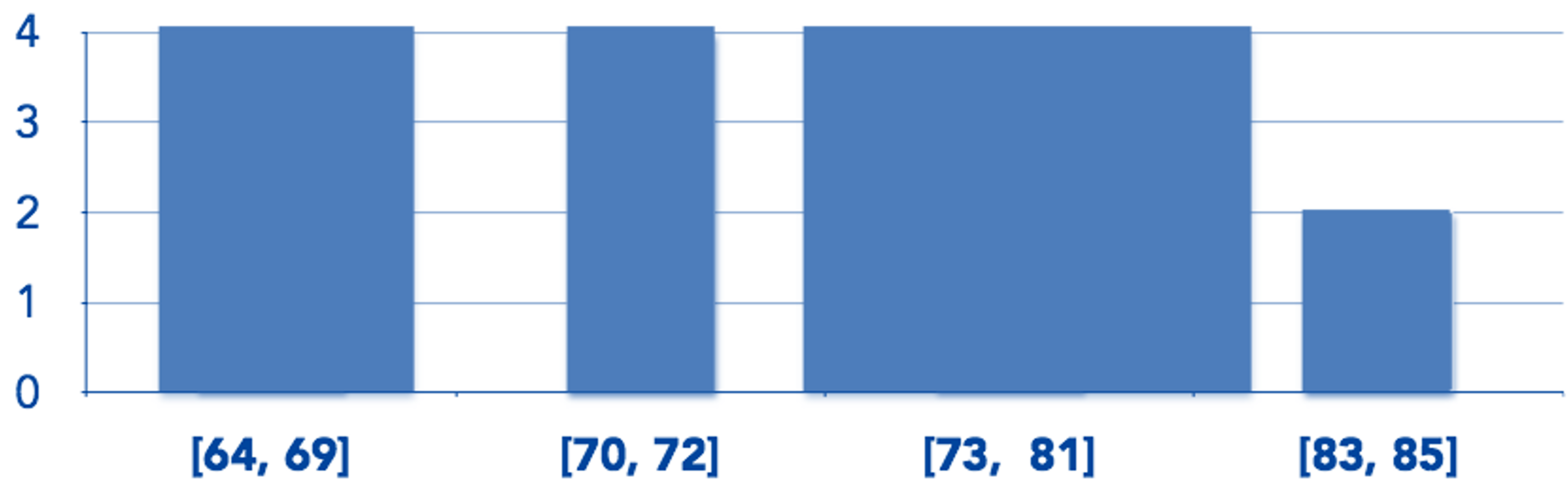
Example: Temperature values are 85, 80, 83, 70, 64, 65, 68, 71, 69, 72, 75, 75, 81,72

- Max=85, Min=64, N=7, W = (85-64)/7=3



Discretisation: Equal-Depth/Frequency/Quantile

- Divides the range [Max, Min] into N intervals
- Each interval including approximately same number of instances
- KBinsDiscretizer(n_bin=4, encode='ordinal', strategy='quantile')
- Example:
  - Sort the 14 Temperature values
  - 64, 65, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85
  - N=4



Missing Values

- Values for one or more variables are missing from recorded observations
- Missing data is a common issue in almost every real dataset
- Caused by varied factors:
  - high cost involved in measuring variables
  - failure of sensors
  - reluctance of respondents in answering certain questions or
  - an ill-designed questionnaire

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	?	No
2	? NaN	?	100K	No
3	No	Single	70K	No
4	Yes	Married	?	No
5	No	?	95K	Yes
6	?	Married	60K	No
7	Yes	Divorced	220K	No
8	No	?	?	Yes
9	?	Married	75K	No
10	No	Single	90K	Yes

Type of Missing Data

Three types of missing data:

- Missing completely at random (MCAR)
  - missing is unrelated to the variable of interests and other variables
  - e.g. survey responses are missing due to occasional data entry errors -> unrelated to respondents or survey questions
- Missing at random (MAR)
  - missing depends on other observed variables but not on the value of the missing data itself
  - e.g. if the likelihood of missing income data in a survey depends on the respondent's education level, but not on the actual income itself
- Missing not at random (MNAR)
  - missing depends on both other observed variables and the missing data itself
  - e.g. high-income individuals are less likely to disclose their income -> missingness is higher for individuals with higher actual incomes

Handling missing values

- Deletion approaches
  - Omits all records containing missing values. Only applies:
    - Missing data introduced in the MCAR mode,
    - When data contains less than 5% of missing values

- Imputation (estimation) approaches
  - Fill missing values with plausible values
  - Mean/Mode imputation
  - KNN imputation

Delete Imcomplete Data Observations

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	?	Married	?	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	?	Divorced	?	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	?	?	75K	No
10	No	Single	90K	Yes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
10	No	Single	90K	Yes

Delete Data Attributes with Missing Values

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	?	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	?	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	?	Married	75K	No
10	No	Single	90K	Yes

Tid		Marital Status	Taxable Income	Cheat
1		Single	125K	No
2		Married	100K	No
3		Single	70K	No
4		Married	120K	No
5		Divorced	95K	Yes
6		Married	60K	No
7		Divorced	220K	No
8		Single	85K	Yes
9		Married	75K	No
10		Single	90K	Yes

Imputation Approaches

- Mean imputation: for continuous attributes
  - Fills with average complete values
  - *sklearn.impute.SimpleImputer (strategy='mean')*
- Mode imputation: for categorical attributes
  - Fills with the most frequent value
  - *sklearn.impute.SimpleImputer (strategy='most_frequent')*
- KNN imputation
  - Find K nearest neighbours using observed values
  - Estimate the missing value by the mean/mode from the K neighbours
  - *sklearn.impute.KNNImputer()*

Estimate Missing Values

Tid	Refund	Marital Status	Taxable Income	Cheat
1	?	Single	125K	No
2	No	?	?	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	?	Yes
6	No	?	60K	No
7	Yes	Divorced	220K	No
8	?	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
	No	**Single**	**105K**	

most common/ mean value

Tid	Refund	Marital Status	Taxable Income	Cheat
1	**No**	Single	125K	No
2	No	**Single**	**105K**	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	**105K**	Yes
6	No	**Single**	60K	No
7	Yes	Divorced	220K	No
8	**No**	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

KNN imputation

Tid	Refund	Marital Status	Taxable Income	Cheat
1	?	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	?	Divorced	95K	Yes
6	?	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	?	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	?	Divorced	95K	Yes
6	?	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

K-NN

Tid	Refund	Marital Status	Taxable Income	Cheat
1	**No**	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	**No**	Divorced	95K	Yes
6	**No**	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Imputation Approaches

- Mean/Mode imputation:
  - (+) Simple, fast
  - (+) Doesn't change the mean/mode of attributes
  - (-) Loss of information, depends on data types

- KNN imputation
  - (+) Capture complex relationship
  - (+) Flexible
  - (-) High computational complexity, Parameters

Summary

- Data preprocessing is an important step in KDD/DM
- Encoding categorical data
- Data normalisation
- Data discretisation
- Missing data