DATA PREPARATION AND VISUALIZATION

Department of Mathematical Economics

National Economics University
https://www.neu.edu.vn/

Introduction

Learning Outcomes

- The importance of data preparation for predictive modeling machine learning projects
- How to prepare data in a way that avoids data leakage, and in turn, incorrect model evaluation
- How to identify and handle problems with messy data, such as outliers and missing values
- How to identify and remove irrelevant and redundant input variables with feature selection methods

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Learning Outcomes

- How to know which feature selection method to choose based on the data types of the variable
- How to scale the range of input variables using normalization and standardization techniques
- How to encode categorical variables as numbers and numeric variables as categories
- How to transform the probability distribution of input variables
- How to transform a dataset with different variable types and how to transform target variables
- Making informative visualizations using matplotlib and seaborn packages

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Course Details

- **Part I: Data Wrangling with Pandas**
  - Chapter 1: Getting Started with Pandas
    - Introduction to pandas Data Structures
    - Essential Functionality
    - Summarizing and Computing Descriptive Statistics
  - Chapter 2: Data Aggregation and Group Operations
    - Groupby Mechanics
    - Data Aggregation
    - Apply: General split-apply-combine
    - Pivot Tables and Cross-Tabulation

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Course Details

- **Part I: Data Wrangling with Pandas**
  - Chapter 3: Advanced pandas
    - Categorical Type in pandas
    - Computations with Categoricals
    - Categorical Methods
    - Grouped Time Resampling

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Course Details

- **Part II: Data Cleaning and Preparation**
  - Chapter 4: Data Cleaning
    - Handling Missing Data
    - Basic Data Cleaning: Delete Columns, Remove Rows or Columns that contain a single value…
    - Outlier Identification and Removal
  - Chapter 5: String Manipulation
    - Regular Expression Basic
    - Regular Expression Advanced
  - Chapter 6: Working with Dates and Times in Python

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Course Details

- **Part II: Data Cleaning and Preparation**
  - Chapter 7: Data Transforms
    - How to Scale Numerical Data
    - How to Scale Data with Outliers
    - How to Encode Categorical Data
    - How to Make Distributions More Gaussian
    - How to change numerical data distributions
    - How to transform numerical to categorical data
    - How to derive new input variables
    - How to save and load data transforms
  - Chapter 8: Feature Selection
    - How to Select Categorical Input Features
    - How to Select Numerical Input Features
    - How to Use Feature Importance

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Course Details

- **Part II: Data Cleaning and Preparation**
  - Chapter 9: Ploting and Visualization
    - A brief matplotlib API Primer
    - Plotting with pandas and seaborn
    - Other Python Visualization Tools
  - Chapter 10: Time Series
    - Time Series Basics
    - Date Ranges, Frequencies, and Shifting
    - Time Zone Handling
    - Resampling and Frequency Conversion
    - Moving Window Functions

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Data Preparation in a Machine Learning Project

What is Data Preparation

- On a predictive modeling project, such as classification or regression, raw data typically cannot be used directly. This is because of reasons such as:
  - Machine learning algorithms require data to be numbers
  - Some machine learning algorithms impose requirements on the data
  - Statistical noise and errors in the data may need to be corrected
  - Complex nonlinear relationships may be teased out of the data
- We can define data preparation as the transformation of raw data into a form that is more suitable for modeling

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

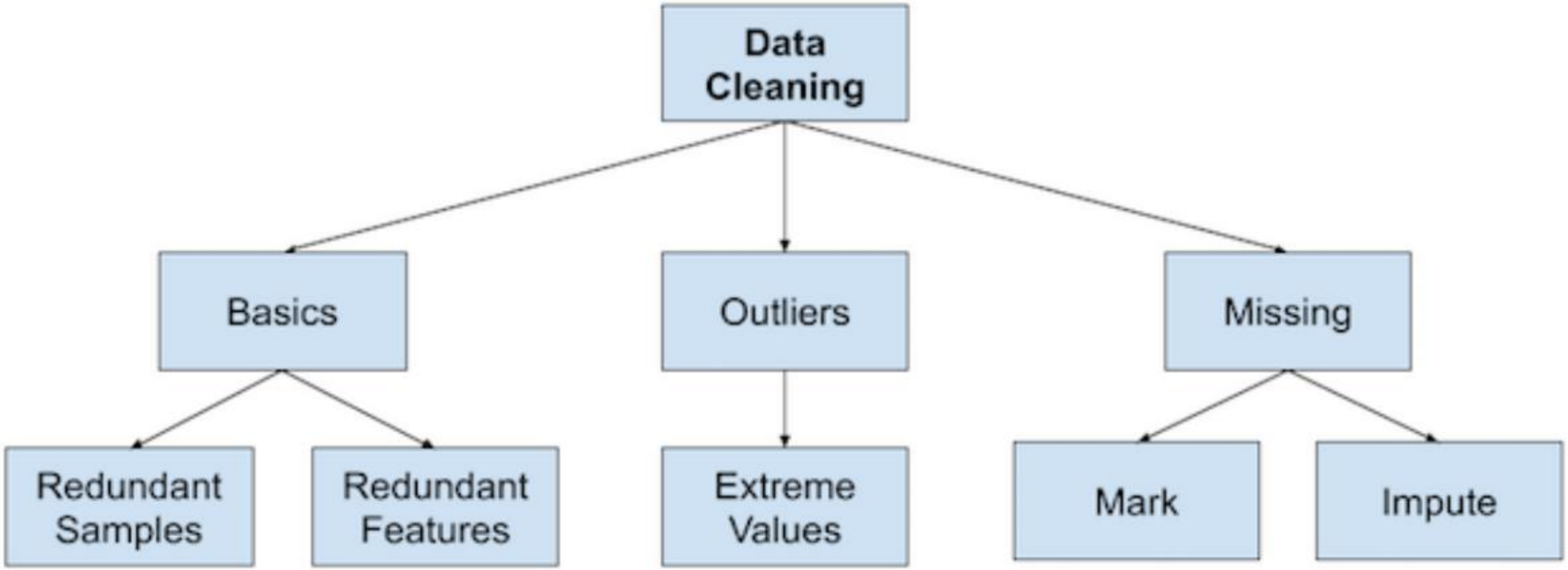
What is Data Preparation

- The common tasks
  - Data Cleaning: Identifying and correcting mistakes or errors in the data
  - Feature Selection: Identifying those input variables that are most relevant to the task
  - Data Transformations: Changing the scale or distribution of variables
  - Feature Engineering: Deriving new variables from available data
  - Dimensionality Reduction: Creating compact projections of the data

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Data Cleaning

Overview of Data Cleaning

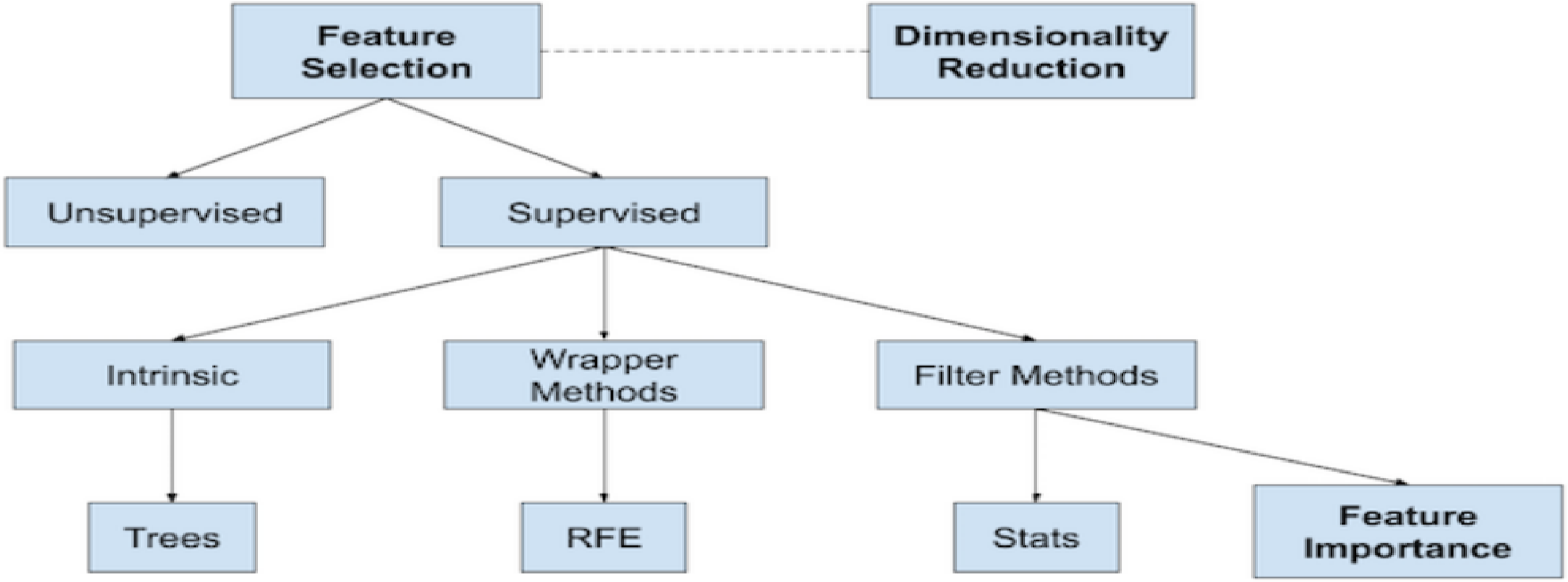


Copyright@2021 by Faculty of Economic Mathematics
National Economics University

Feature Selection

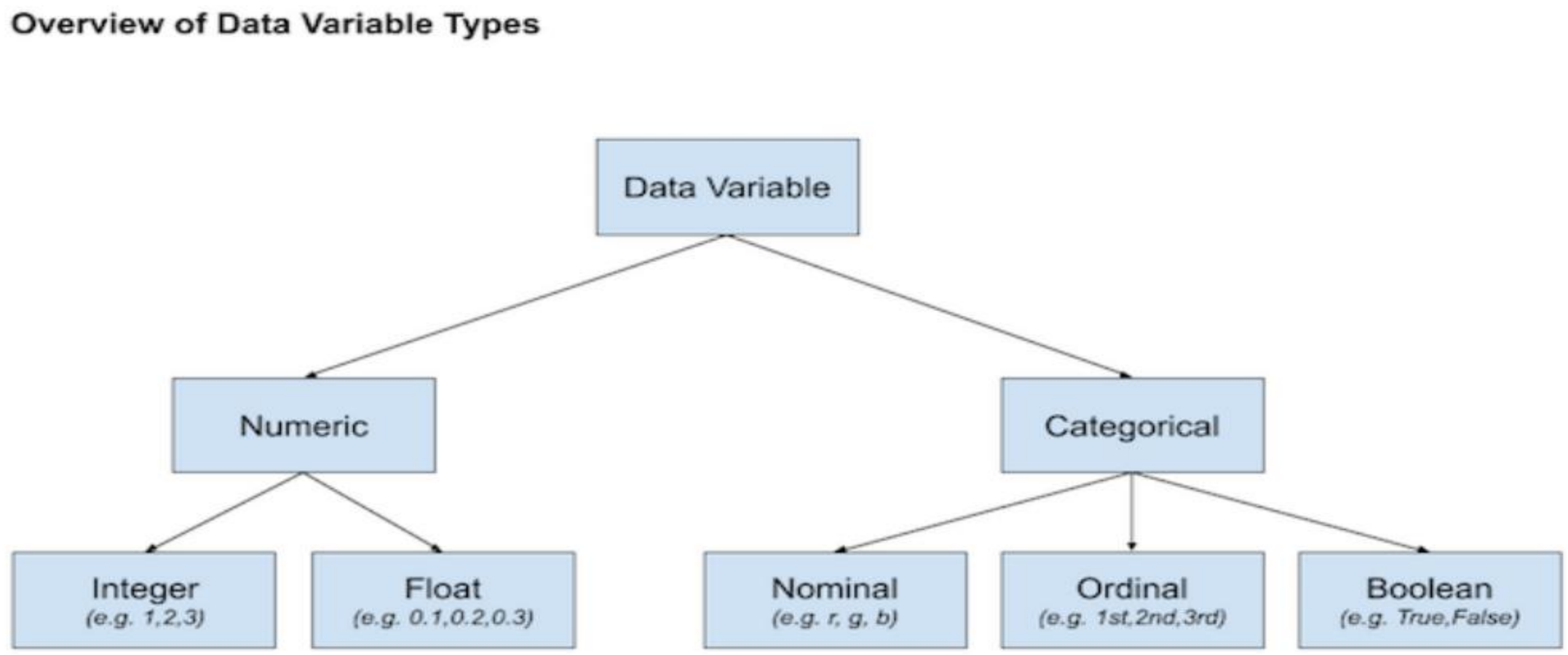
Feature selection refers to techniques for selecting a subset of input features that are most relevant to the target variable that is being predicted

Overview of Feature Selection Techniques



Copyright@2021 by Faculty of Economic Mathematics
National Economics University

Data Transformation



Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Data Transformation

- **Discretization Transform**: Encode a numeric variable as an ordinal variable
- **Ordinal Transform**: Encode a categorical variable into an integer variable
- **One Hot Transform**: Encode a categorical variable into binary variables
- **Normalization Transform**: Scale a variable to the range 0 and 1
- **Standardization Transform**: Scale a variable to a standard Gaussian
- **Power Transform**: Change the distribution of a variable to be more Gaussian
- **Quantile Transform**: Impose a probability distribution such as uniform or Gaussian

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Data Transformation

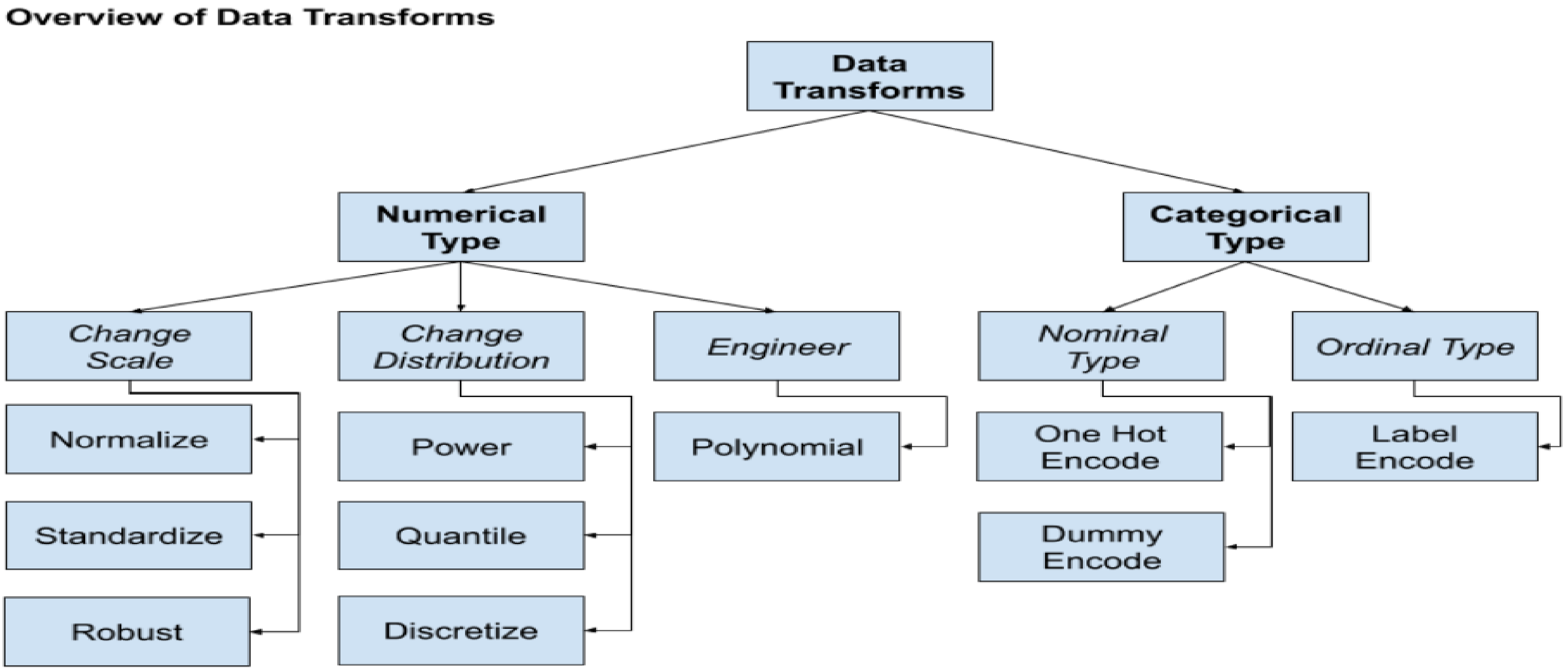


Figure 3.4: Overview of Data Transform Techniques.

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Feature Engineering

- Feature engineering refers to the process of creating new input variables from the available data
  - Adding a Boolean flag variable for some state
  - Adding a group or global summary statistic, such as a mean
  - Adding new variables for each component of a compound variable, such as a date-time
  - **Polynomial Transformation**: Create copies of numerical input variables that are raised to a power

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Dimensionality Reduction

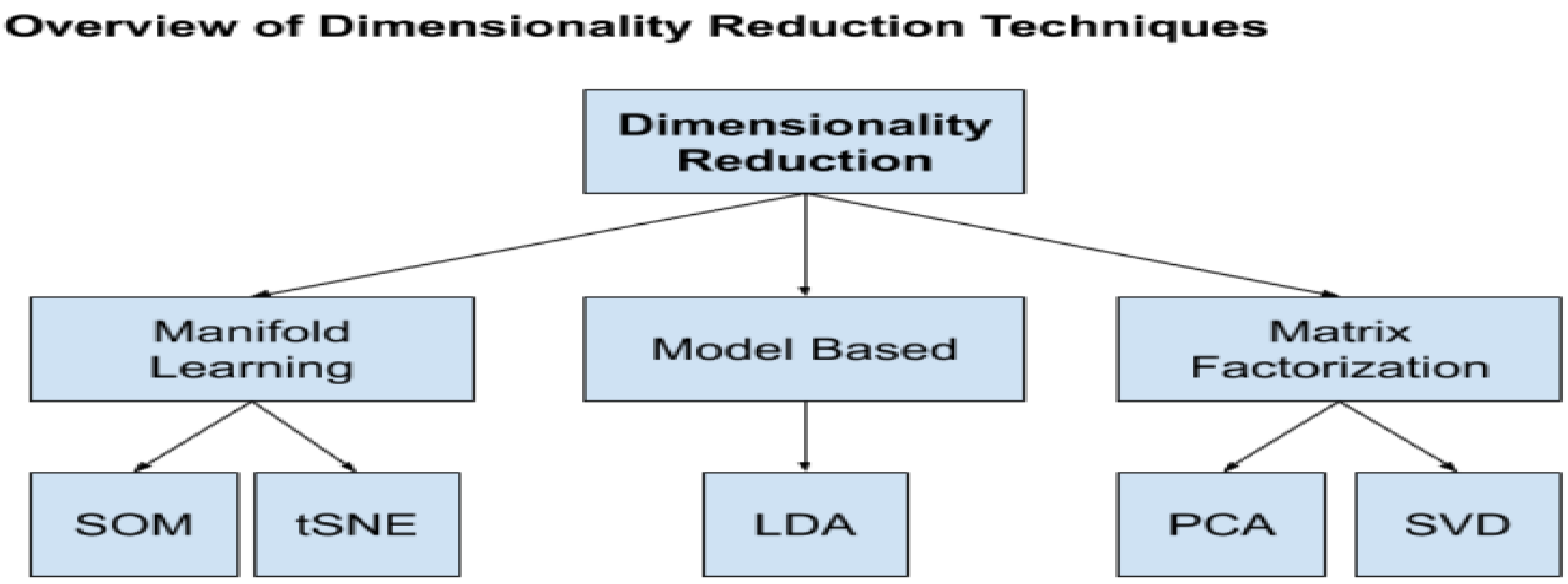


Figure 3.5: Overview of Dimensionality Reduction Techniques.

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Data Preparation Without Data Leakage

- **Data leakage** refers to a problem where information about the holdout dataset, such as a test or validation dataset, is made available to the model in the training dataset
  - Ex: Using MinMaxScaler for whole dataset
- Applying preprocessing techniques to the entire dataset will cause the model to learn not only the training set but also the test set. The test set should be new and previously unseen for any model.
- Data leakage will likely result in an incorrect estimate of a model's performance on the problem
- The problem with applying data preparation techniques before splitting data for model evaluation is that it can lead to data leakage
- Data preparation must be fit on the training dataset only to avoid data leakage

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University

Train-Test Evaluation With Naive Data Preparation

```
# naive approach to normalizing the data before splitting the data and evaluating the model
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
# define dataset
X, y = make_classification(n_samples=1000, n_features=20, n_informative=15, n_redundant=5,
    random_state=7)
# standardize the dataset
scaler = MinMaxScaler()
X = scaler.fit_transform(X)
# split into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=1)
# fit the model
model = LogisticRegression()
model.fit(X_train, y_train)
# evaluate the model
yhat = model.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, yhat)
print('Accuracy: %.3f' % (accuracy*100))
```

Accuray: 84.848%

Copyright@2021 by Faculty of Economic Mathematics
National Economics University

Train-Test Evaluation With Correct Data Preparation

```
# correct approach for normalizing the data after the data is split before the model is
    evaluated
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
# define dataset
X, y = make_classification(n_samples=1000, n_features=20, n_informative=15, n_redundant=5,
    random_state=7)
# split into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=1)
# define the scaler
scaler = MinMaxScaler()
# fit on the training dataset
scaler.fit(X_train)
# scale the training dataset
X_train = scaler.transform(X_train)
# scale the test dataset
X_test = scaler.transform(X_test)
# fit the model
model = LogisticRegression()
model.fit(X_train, y_train)
# evaluate the model
yhat = model.predict(X_test)
# evaluate predictions
accuracy = accuracy_score(y_test, yhat)
print('Accuracy: %.3f' % (accuracy*100))
```

Accuracy: 85.455%

Copyright@2021 by **Faculty of Economic Mathematics**
National Economics University