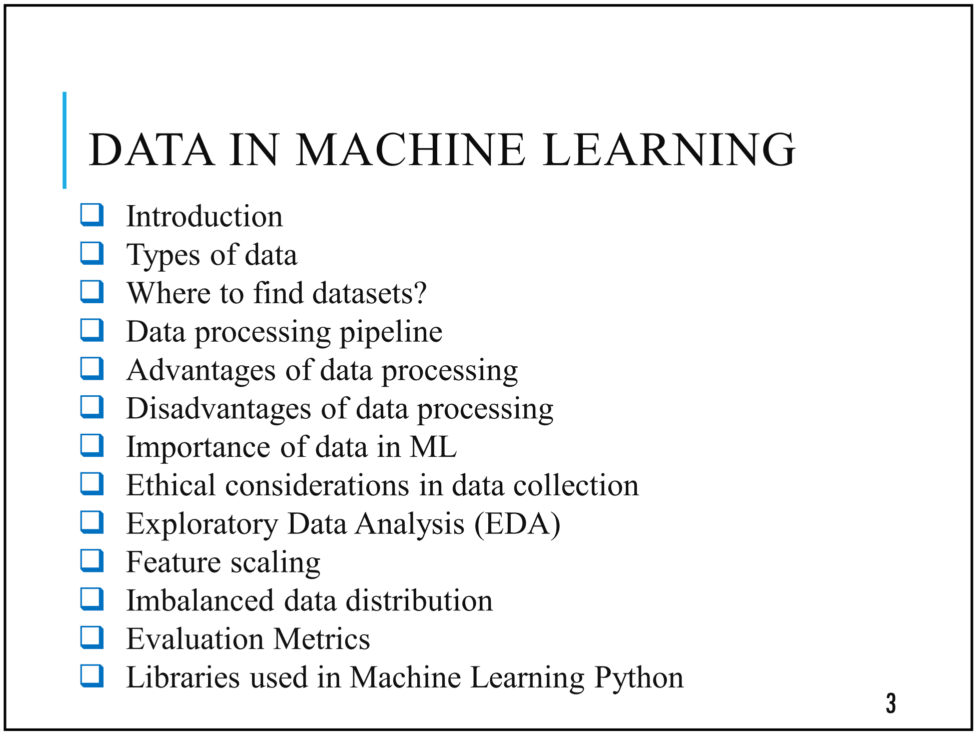
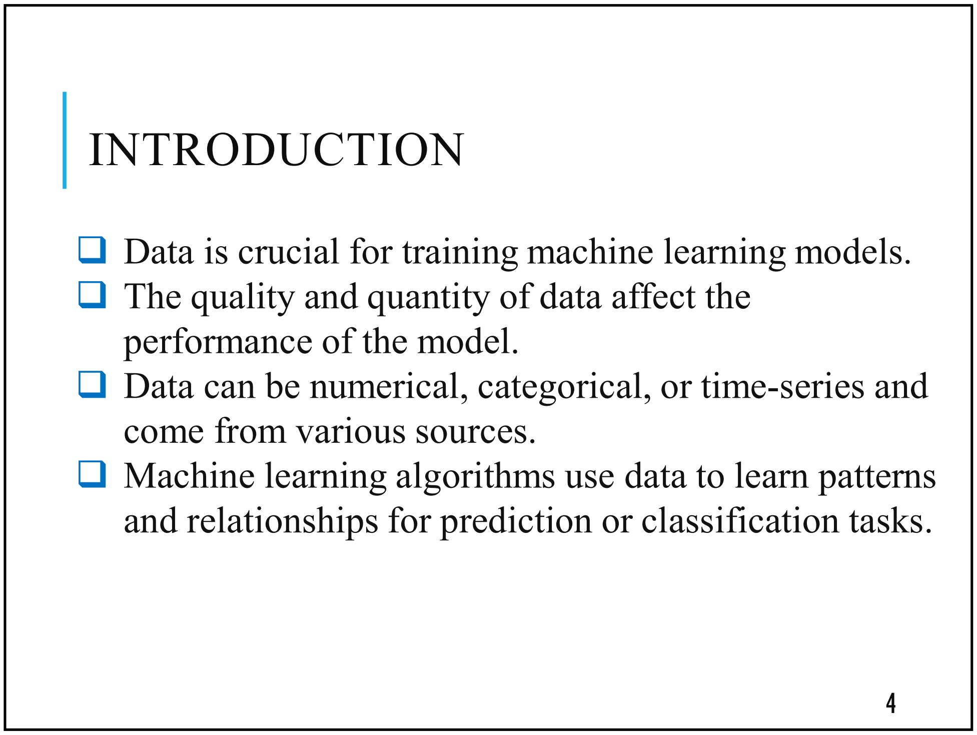
ELEC8900-57 SPECIAL TOPICS: MACHINE LEARNING

DATA IN MACHINE LEARNING



DATA IN MACHINE LEARNING

- Introduction
- Types of data
- Where to find datasets?
- Data processing pipeline
- Advantages of data processing
- Disadvantages of data processing
- Importance of data in ML
- Ethical considerations in data collection
- Exploratory Data Analysis (EDA)
- Feature scaling
- Imbalanced data distribution
- Evaluation Metrics
- Libraries used in Machine Learning Python



INTRODUCTION

- Data is crucial for training machine learning models.
- The quality and quantity of data affect the performance of the model.
- Data can be numerical, categorical, or time-series and come from various sources.
- Machine learning algorithms use data to learn patterns and relationships for prediction or classification tasks.

INTRODUCTION

❑ Data is unprocessed facts, values, text, sound, or pictures that have not been interpreted or analyzed.

❑ Data can be

1. labeled
2. unlabeled,

❑ Data is typically divided into training and testing sets.

❑ Data preprocessing, including:

- ❖ cleaning
- ❖ normalizing
- ❖ handling missing values
- ❖ ….

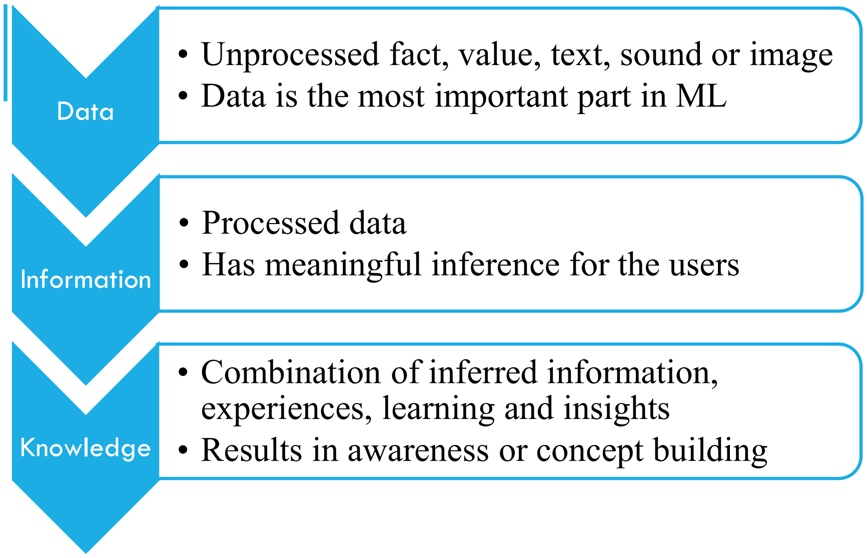
INTRODUCTION

❑ Enterprises spend a lot of money to gather data.

❑ An example is Facebook acquiring WhatsApp for $19 billion to access user information.

❑ Information is data that has been interpreted and manipulated to have meaningful inference for users.

❑ Knowledge is a combination of inferred information, experiences, learning, and insights that results in awareness or concept building.



Data

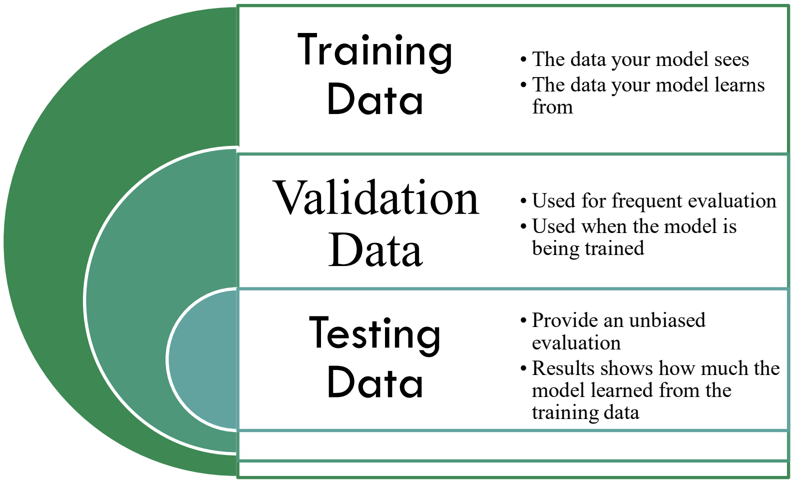
- Unprocessed fact, value, text, sound or image
- Data is the most important part in ML

Information

- Processed data
- Has meaningful inference for the users

Knowledge

- Combination of inferred information, experiences, learning and insights
- Results in awareness or concept building



Training Data

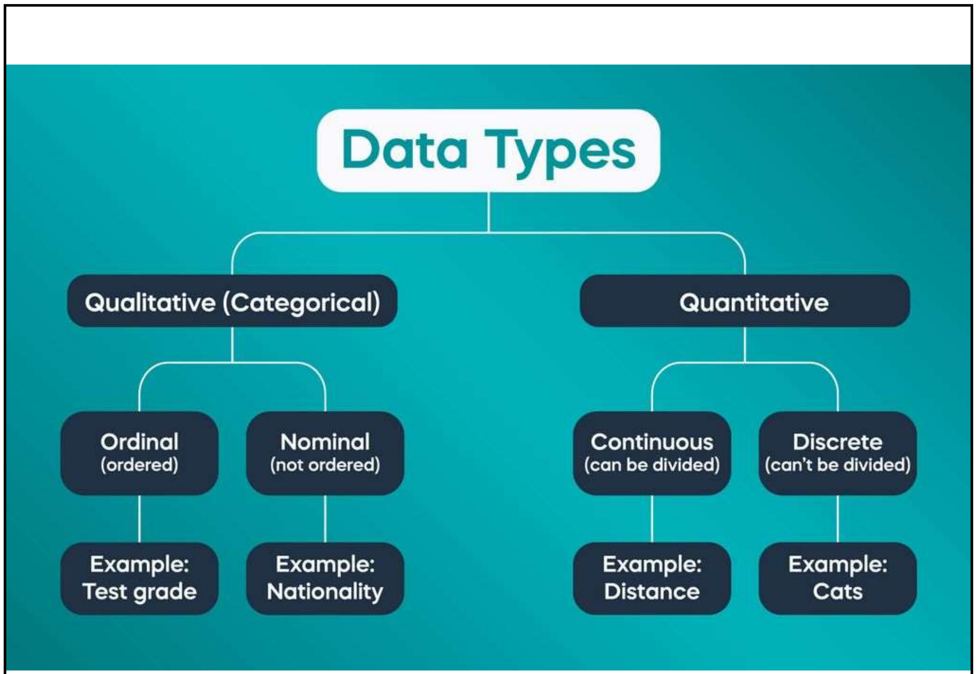
- The data your model sees
- The data your model learns from

Validation Data

- Used for frequent evaluation
- Used when the model is being trained

Testing Data

- Provide an unbiased evaluation
- Results shows how much the model learned from the training data



Data Types

Qualitative (Categorical)

Quantitative

Ordinal (ordered)

Nominal (not ordered)

Continuous (can be divided)

Discrete (can't be divided)

Example: Test grade

Example: Nationality

Example: Distance

Example: Cats

[Source: https://365datascience.com/resources/blog/thumb@1024_xzk3rl06exk-data-types.jpeg]

Where to Search for Datasets?

❑ Kaggle- https://www.kaggle.com/
This data science site contains a diverse set of compelling, independently-contributed datasets for machine learning. If you're looking for niche datasets, Kaggle's search engine allows you to specify categories to ensure the datasets you find will fit your bill.

❑ Google Dataset Search: https://datasetsearch.research.google.com/
Dataset Search contains over 25 million datasets from all across the web. Whether they're hosted on a publisher's site, a government domain, or a researcher's blog, Dataset Search can find it.

Where to Search for Datasets?

AWS Open Data Registry- https://registry.opendata.aws/

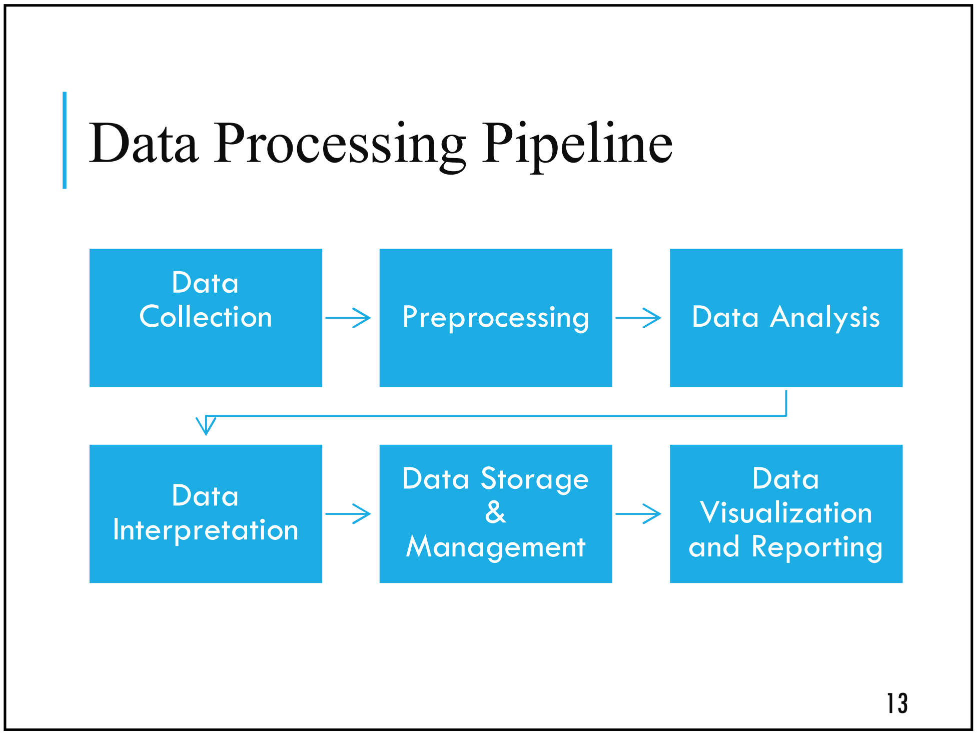
- ❑ Amazon has their hands in the open dataset cookie jar as well. The shopping juggernaut brings their trademark resourcefulness to the dataset searching game.
- ❑ One key perk that differentiates AWS Open Data Registry is its user feedback feature, which allows users to add and modify datasets.
- ❑ Experience with AWS is also highly preferred in the job marketplace.

Where to Search for Datasets?

❑ Wikipedia ML Datasets: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

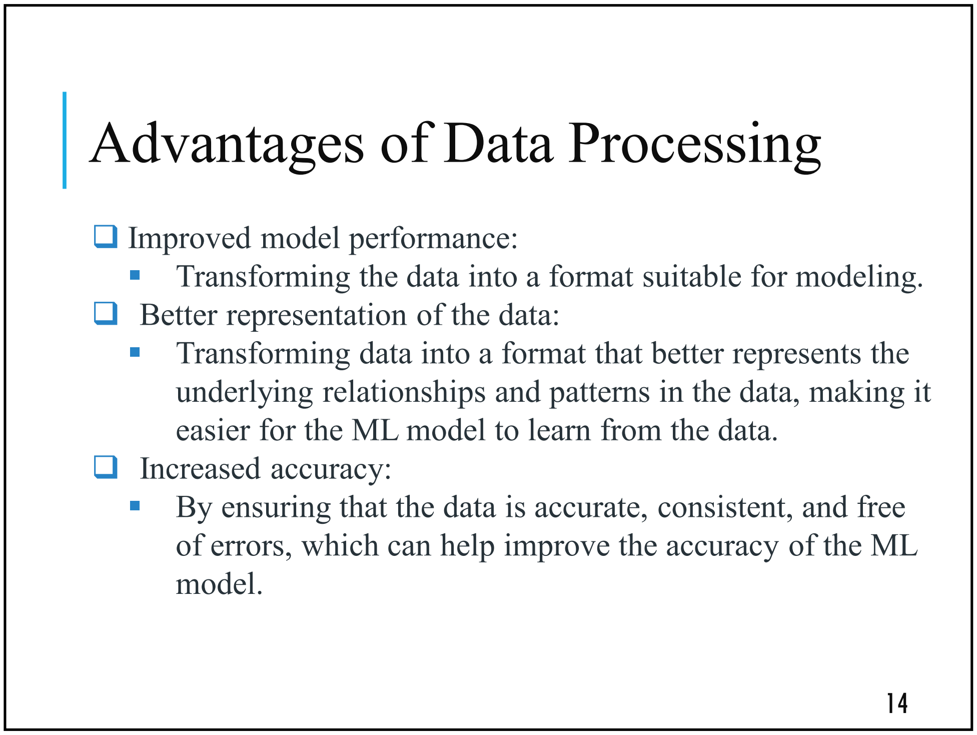
This Wikipedia page features diverse datasets for machine learning including signal, image, sound, and text.

❑ The following website contains links to the best Machine Learning datasets: https://imerit.net/blog/the-60-best-free-datasets-for-machine-learning-all-pbm/



Data Processing Pipeline

Data Collection → Preprocessing → Data Analysis → Data Interpretation → Data Storage & Management → Data Visualization and Reporting



Advantages of Data Processing

- ❑ Improved model performance:
  - Transforming the data into a format suitable for modeling.
- ❑ Better representation of the data:
  - Transforming data into a format that better represents the underlying relationships and patterns in the data, making it easier for the ML model to learn from the data.
- ❑ Increased accuracy:
  - By ensuring that the data is accurate, consistent, and free of errors, which can help improve the accuracy of the ML model.

Disadvantages of Data Processing

- Time-consuming: Data processing can be a time-consuming task, especially for large and complex datasets.
- Error-prone: Data processing can be error-prone, as it involves transforming and cleaning the data, which can result in the loss of important information or the introduction of new errors.
- Limited understanding of the data: Data processing can lead to a limited understanding of the data, as the transformed data may not be representative of the underlying relationships and patterns in the data.

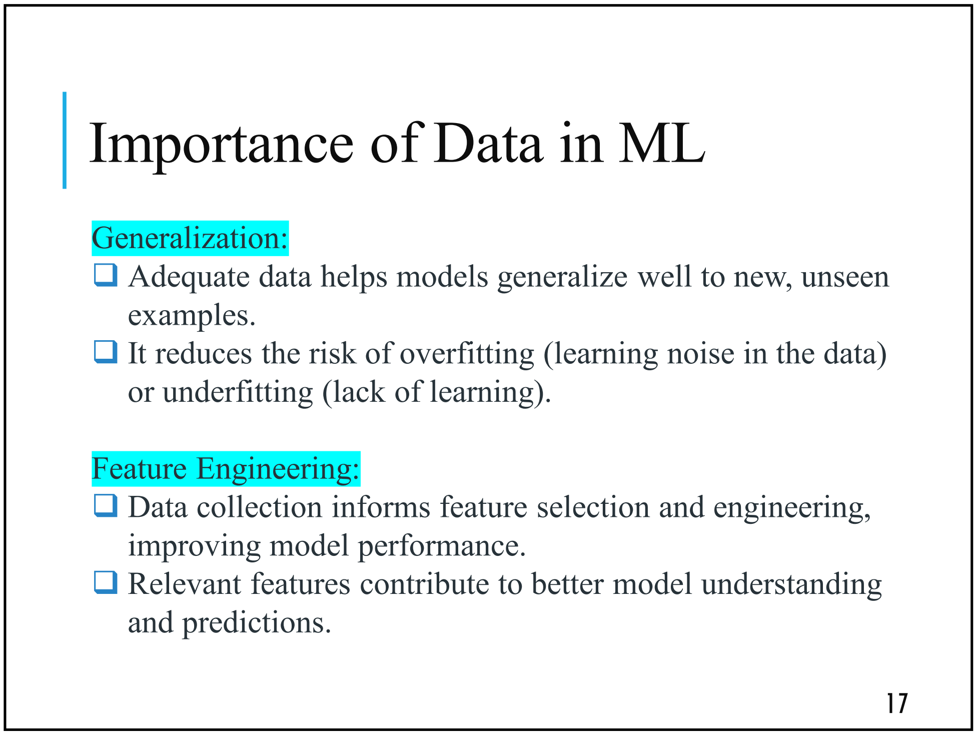
Importance of Data in ML

Foundation of Machine Learning:

- Data is the raw material upon which machine learning models are built.
- It forms the foundation for training, evaluating, and improving models.

Model Training:

- High-quality, diverse data is essential for training accurate and robust machine learning models.
- Models learn patterns and make predictions based on the data they are exposed to.



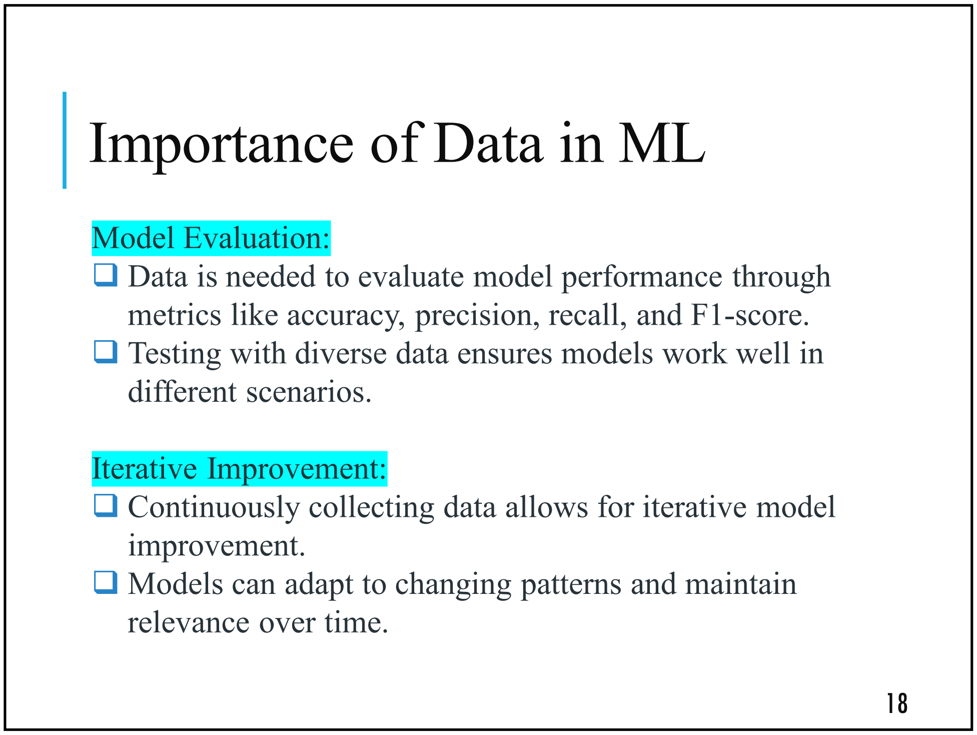
Importance of Data in ML

Generalization:

- ❑ Adequate data helps models generalize well to new, unseen examples.
- ❑ It reduces the risk of overfitting (learning noise in the data) or underfitting (lack of learning).

Feature Engineering:

- ❑ Data collection informs feature selection and engineering, improving model performance.
- ❑ Relevant features contribute to better model understanding and predictions.



Importance of Data in ML

Model Evaluation:

- ❑ Data is needed to evaluate model performance through metrics like accuracy, precision, recall, and F1-score.
- ❑ Testing with diverse data ensures models work well in different scenarios.

Iterative Improvement:

- ❑ Continuously collecting data allows for iterative model improvement.
- ❑ Models can adapt to changing patterns and maintain relevance over time.

Importance of Data in ML

Bias and Fairness:

- Careful data collection helps address bias and fairness concerns in machine learning.
- Biased data can lead to biased models, which can have ethical and social implications.

Data Preprocessing:

- Data collection often involves preprocessing steps like cleaning, normalization, and handling missing values.
- Clean data leads to more reliable model outcomes.

19

Importance of Data in ML

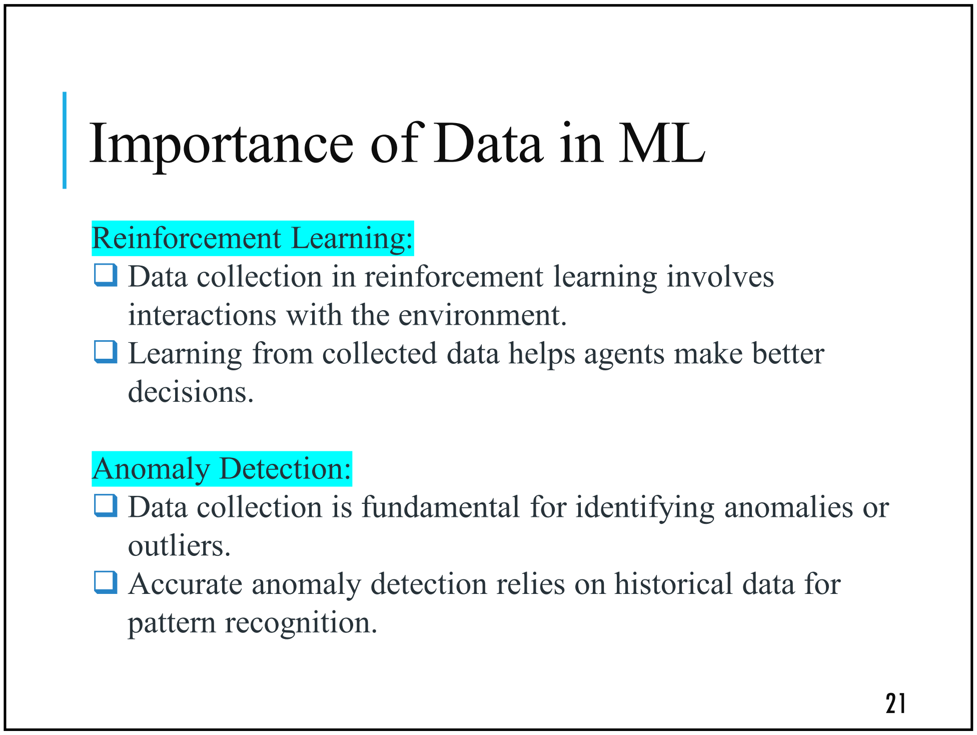
Supervised Learning:

- In supervised learning, labeled data (input-output pairs) is crucial for training models.
- The availability and quality of labeled data impact model accuracy.

Unsupervised Learning:

- Unlabeled data is used in unsupervised learning for clustering and dimensionality reduction.
- Collecting relevant unlabeled data can improve model insights.

20



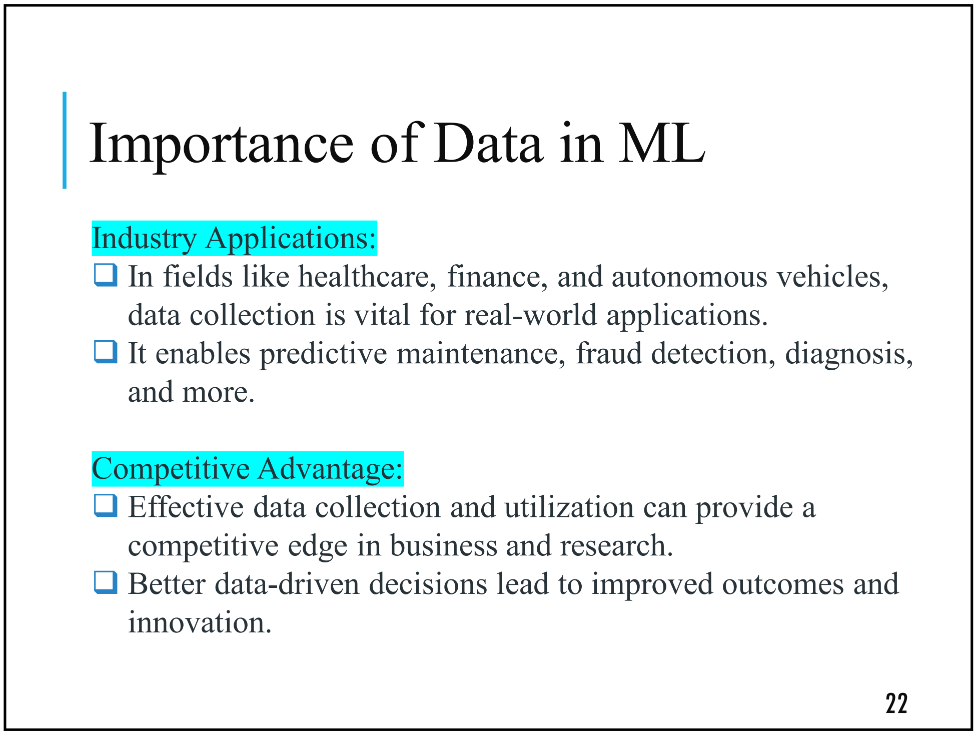
Importance of Data in ML

Reinforcement Learning:

- Data collection in reinforcement learning involves interactions with the environment.
- Learning from collected data helps agents make better decisions.

Anomaly Detection:

- Data collection is fundamental for identifying anomalies or outliers.
- Accurate anomaly detection relies on historical data for pattern recognition.



Importance of Data in ML

Industry Applications:

- In fields like healthcare, finance, and autonomous vehicles, data collection is vital for real-world applications.
- It enables predictive maintenance, fraud detection, diagnosis, and more.

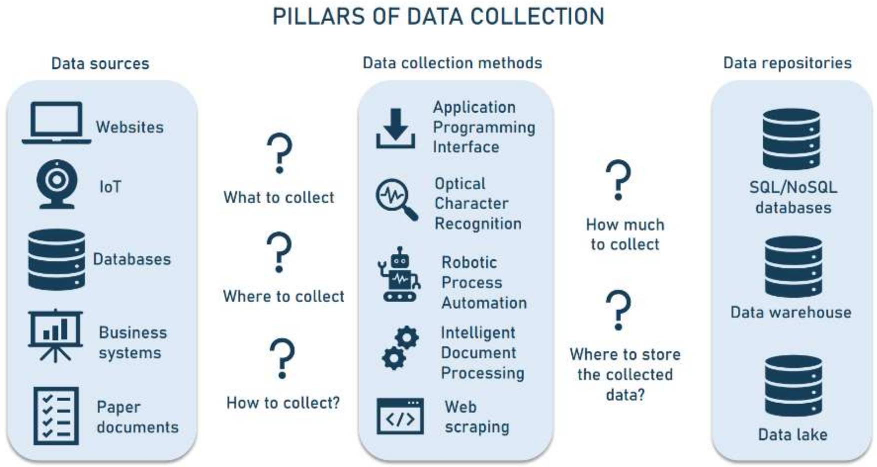
Competitive Advantage:

- Effective data collection and utilization can provide a competitive edge in business and research.
- Better data-driven decisions lead to improved outcomes and innovation.

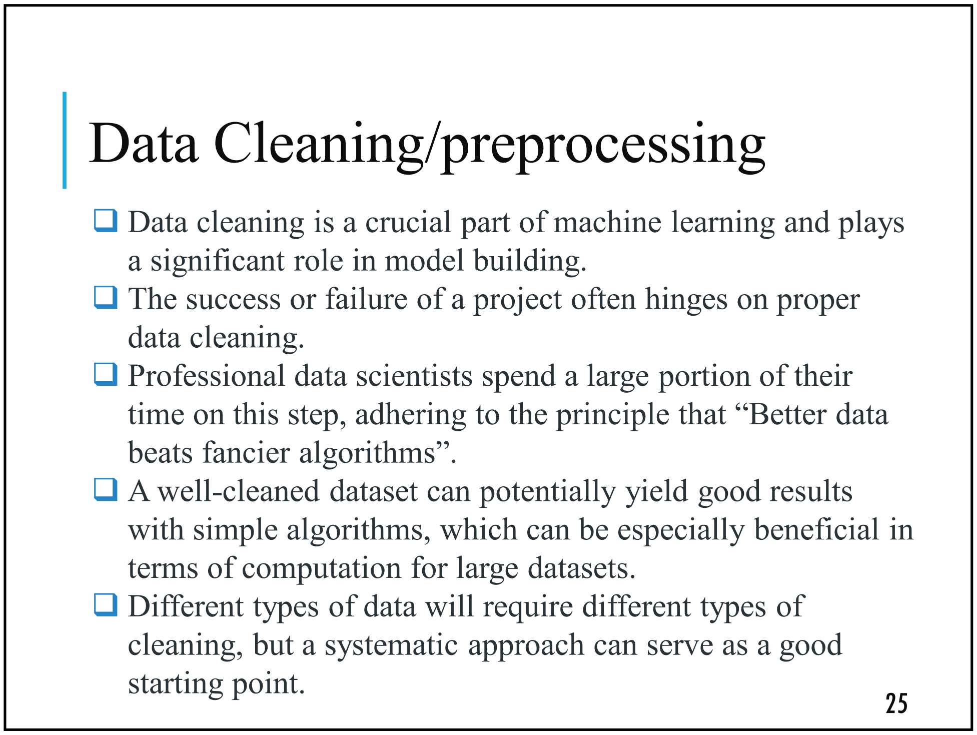
Ethical Considerations in Data Collection

- ❑ Responsible Data Collection: Ensures privacy, complies with regulations, and safeguards user rights through transparency and informed consent.
- ❑ Bias Mitigation and Fairness: Mitigates bias and ensures fairness.
- ❑ Security and Ownership: Secures data, protects privacy, and clarifies ownership.
- ❑ Accountability and Monitoring: Assigns responsibility and conducts continuous monitoring.

DATA COLLECTION

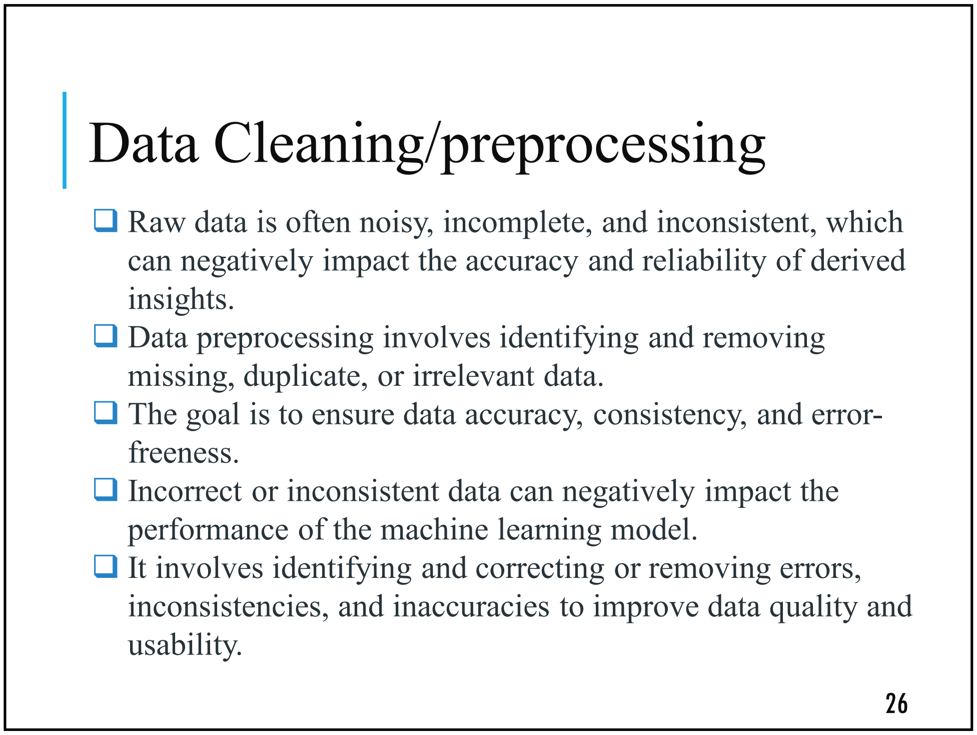


https://www.altexsoft.com/blog/data-collection-machine-learning/



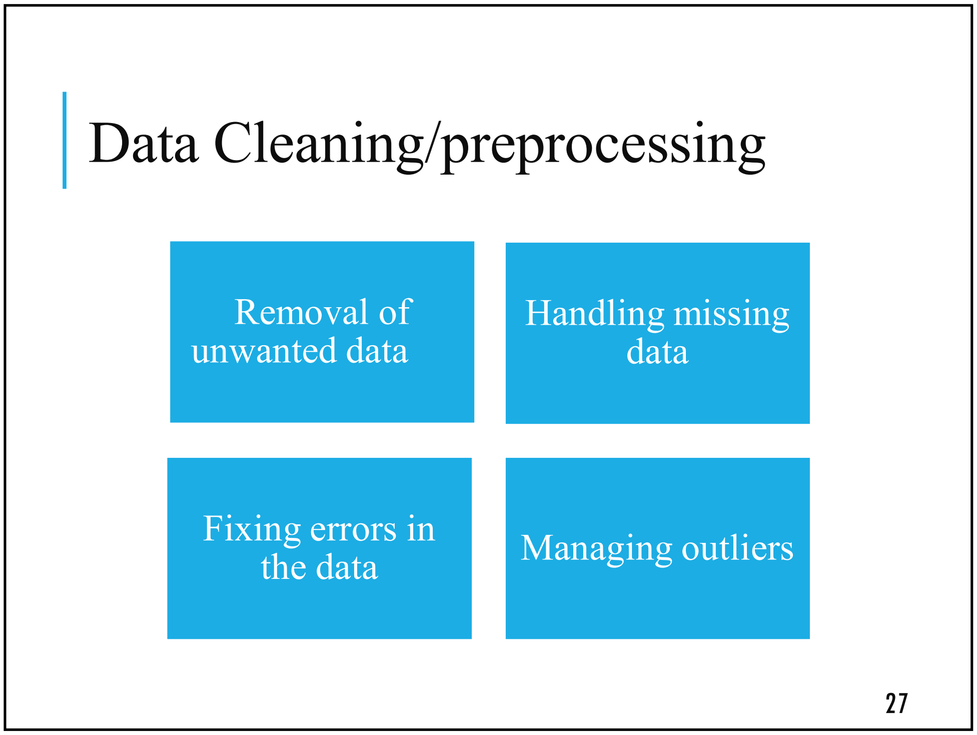
Data Cleaning/preprocessing

- ❑ Data cleaning is a crucial part of machine learning and plays a significant role in model building.
- ❑ The success or failure of a project often hinges on proper data cleaning.
- ❑ Professional data scientists spend a large portion of their time on this step, adhering to the principle that “Better data beats fancier algorithms”.
- ❑ A well-cleaned dataset can potentially yield good results with simple algorithms, which can be especially beneficial in terms of computation for large datasets.
- ❑ Different types of data will require different types of cleaning, but a systematic approach can serve as a good starting point.



Data Cleaning/preprocessing

- ❑ Raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of derived insights.
- ❑ Data preprocessing involves identifying and removing missing, duplicate, or irrelevant data.
- ❑ The goal is to ensure data accuracy, consistency, and error-freeness.
- ❑ Incorrect or inconsistent data can negatively impact the performance of the machine learning model.
- ❑ It involves identifying and correcting or removing errors, inconsistencies, and inaccuracies to improve data quality and usability.



Data Cleaning/preprocessing

Removal of unwanted data

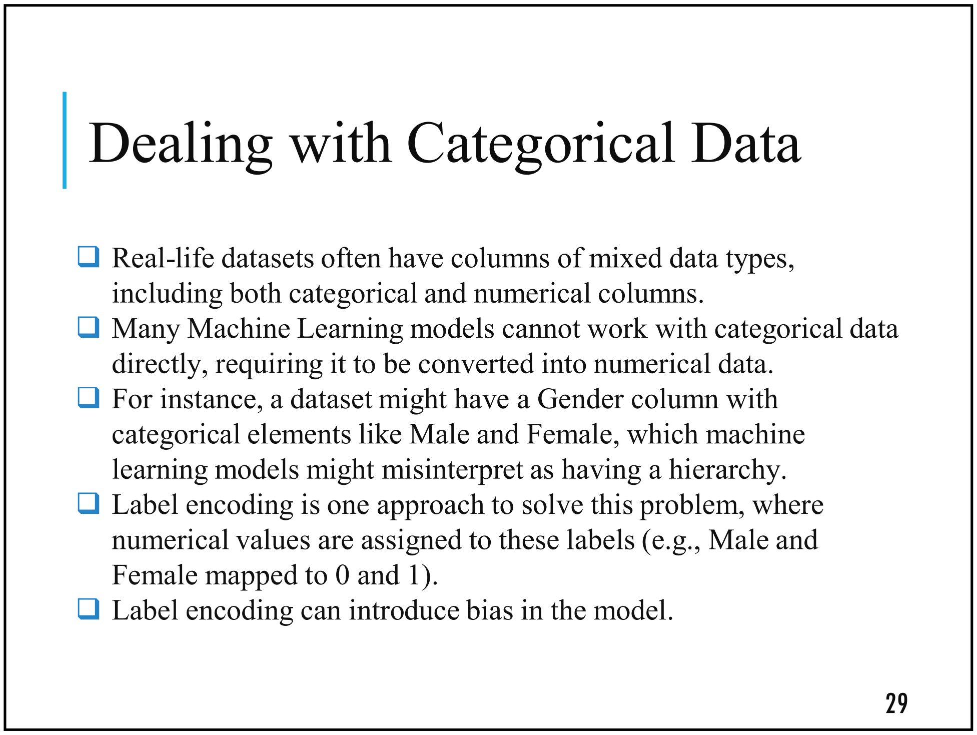
Handling missing data

Fixing errors in the data

Managing outliers

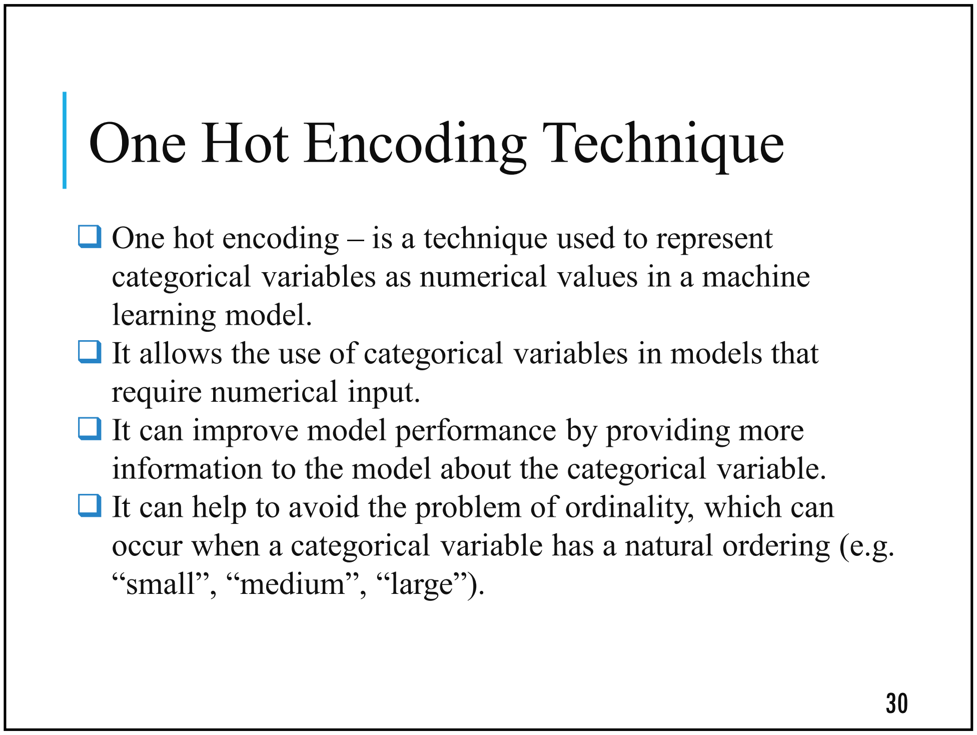
Data Cleaning Example

Reference:
https://www.geeksforgeeks.org/data-cleansing-introduction/
Data:
https://www.kaggle.com/competitions/titanic/data?select=train.csv



Dealing with Categorical Data

- Real-life datasets often have columns of mixed data types, including both categorical and numerical columns.
- Many Machine Learning models cannot work with categorical data directly, requiring it to be converted into numerical data.
- For instance, a dataset might have a Gender column with categorical elements like Male and Female, which machine learning models might misinterpret as having a hierarchy.
- Label encoding is one approach to solve this problem, where numerical values are assigned to these labels (e.g., Male and Female mapped to 0 and 1).
- Label encoding can introduce bias in the model.



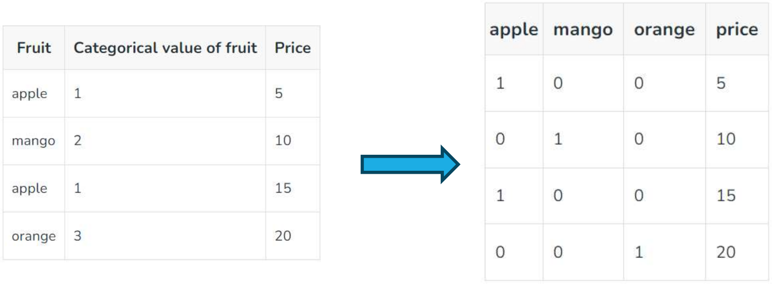
One Hot Encoding Technique

- One hot encoding – is a technique used to represent categorical variables as numerical values in a machine learning model.
- It allows the use of categorical variables in models that require numerical input.
- It can improve model performance by providing more information to the model about the categorical variable.
- It can help to avoid the problem of ordinality, which can occur when a categorical variable has a natural ordering (e.g. “small”, “medium”, “large”).

Disadvantages of One Hot Encoding Technique

- It can lead to increased dimensionality, as a separate column is created for each category in the variable. This can make the model more complex and slow to train.
- It can lead to sparse data, as most observations will have a value of 0 in most of the one-hot encoded columns.
- It can lead to overfitting, especially if there are many categories in the variable and the sample size is relatively small.
- One-hot-encoding is a powerful technique to treat categorical data, but it can lead to increased dimensionality, sparsity, and overfitting. It is important to use it cautiously and consider other methods such as ordinal encoding or binary encoding.

One Hot Encoding Technique



Fruit	Categorical value of fruit	Price
apple	1	5
mango	2	10
apple	1	15
orange	3	20

apple	mango	orange	price
1	0	0	5
0	1	0	10
1	0	0	15
0	0	1	20

Exploratory Data Analysis (EDA)

- EDA refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.
- EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

Goals of EDA

- Data Cleaning
- Descriptive Statistics
- Data Visualization
- Feature Engineering
- Correlation and Relationships
- Data Segmentation
- Hypothesis Generation
- Data Quality Assessment

Feature Scaling

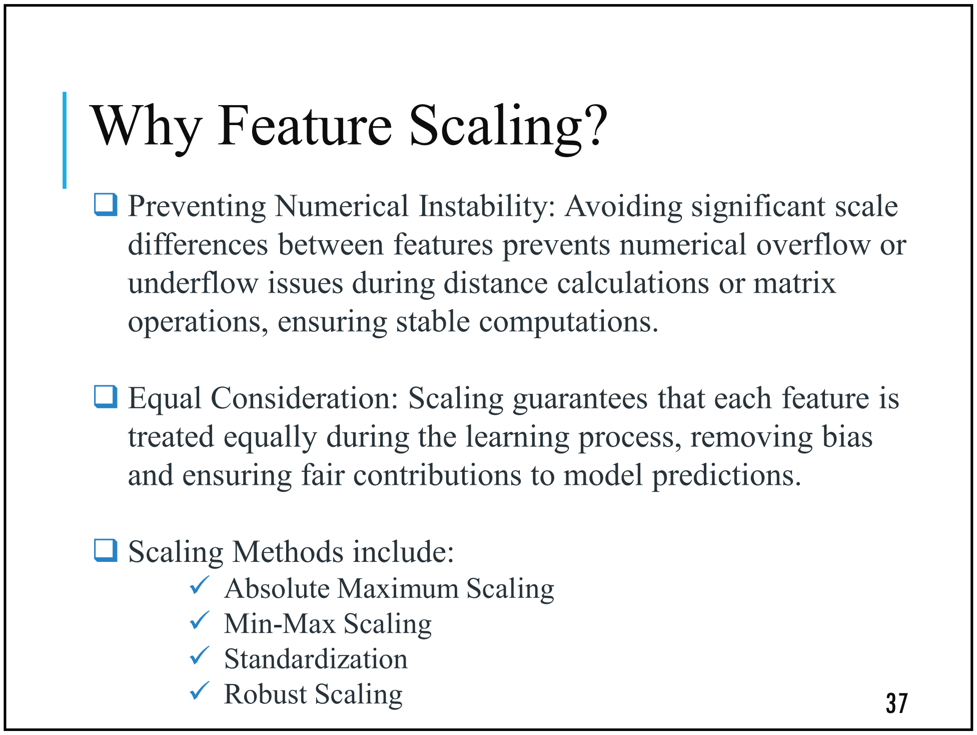
❑ Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing

Why Feature Scaling?

Feature scaling in machine learning serves several purposes:

❑ Comparable Scale and Range: Scaling ensures that all features have similar scales and ranges, preventing larger-scale features from dominating the learning process and ensuring each feature contributes equally.

❑ Algorithm Performance Improvements: Many ML algorithms, like gradient descent, k-nearest neighbors, and support vector machines, perform better or converge faster with scaled features, enhancing overall algorithm performance.



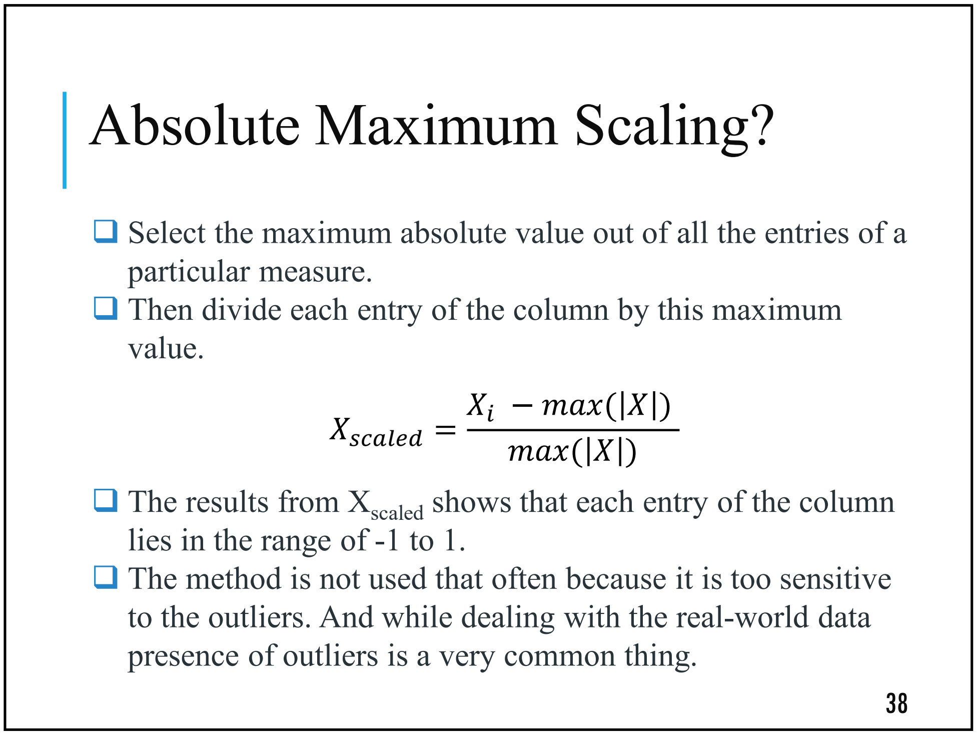
Why Feature Scaling?

- ❑ Preventing Numerical Instability: Avoiding significant scale differences between features prevents numerical overflow or underflow issues during distance calculations or matrix operations, ensuring stable computations.

- ❑ Equal Consideration: Scaling guarantees that each feature is treated equally during the learning process, removing bias and ensuring fair contributions to model predictions.

- ❑ Scaling Methods include:
  - ✓ Absolute Maximum Scaling
  - ✓ Min-Max Scaling
  - ✓ Standardization
  - ✓ Robust Scaling

37



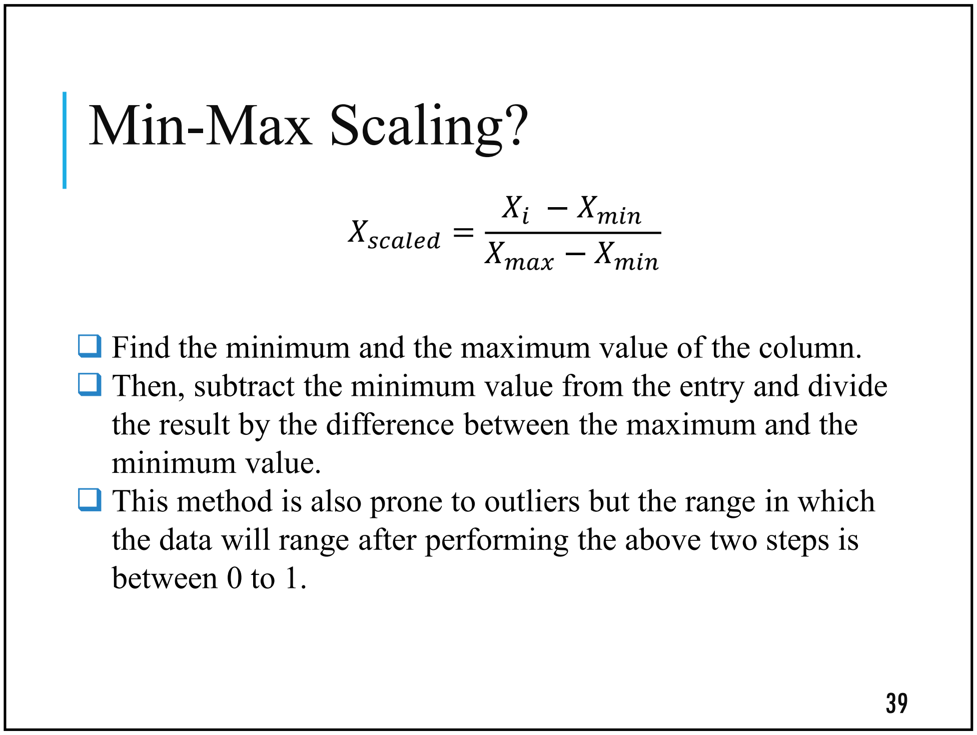
Absolute Maximum Scaling?

- ❑ Select the maximum absolute value out of all the entries of a particular measure.
- ❑ Then divide each entry of the column by this maximum value.

$$X_{scaled} = \frac{X_i \; - max(|X|)}{max(|X|)}$$

- ❑ The results from X_{scaled} shows that each entry of the column lies in the range of -1 to 1.
- ❑ The method is not used that often because it is too sensitive to the outliers. And while dealing with the real-world data presence of outliers is a very common thing.

38



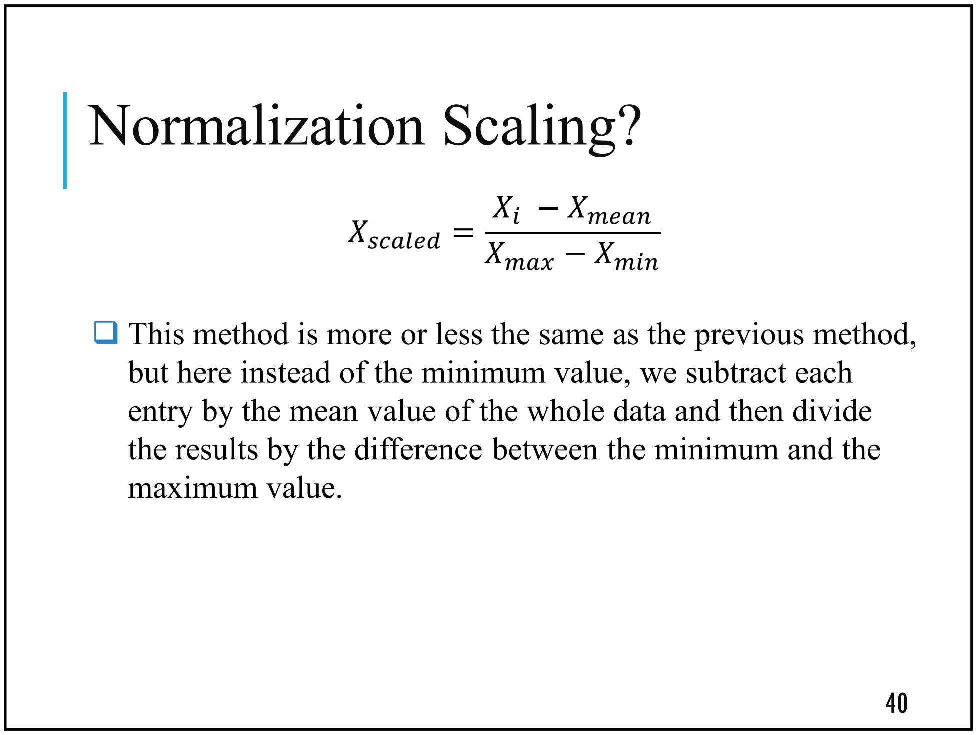
Min-Max Scaling?

$$X_{scaled} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

❑ Find the minimum and the maximum value of the column.

❑ Then, subtract the minimum value from the entry and divide the result by the difference between the maximum and the minimum value.

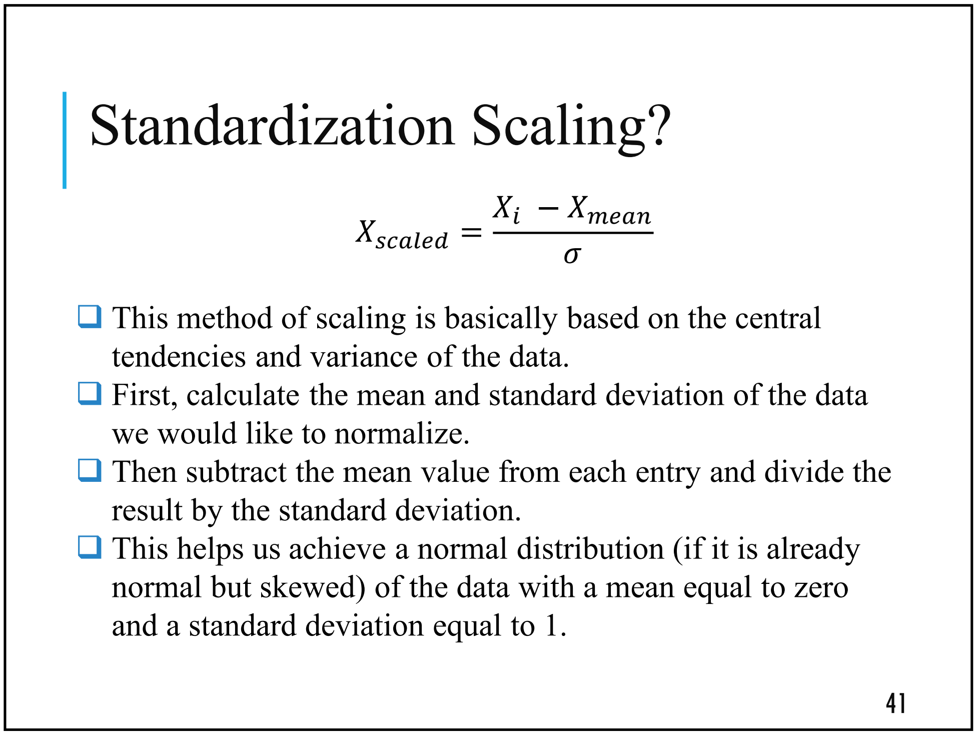
❑ This method is also prone to outliers but the range in which the data will range after performing the above two steps is between 0 to 1.



Normalization Scaling?

$$X_{scaled} = \frac{X_i - X_{mean}}{X_{max} - X_{min}}$$

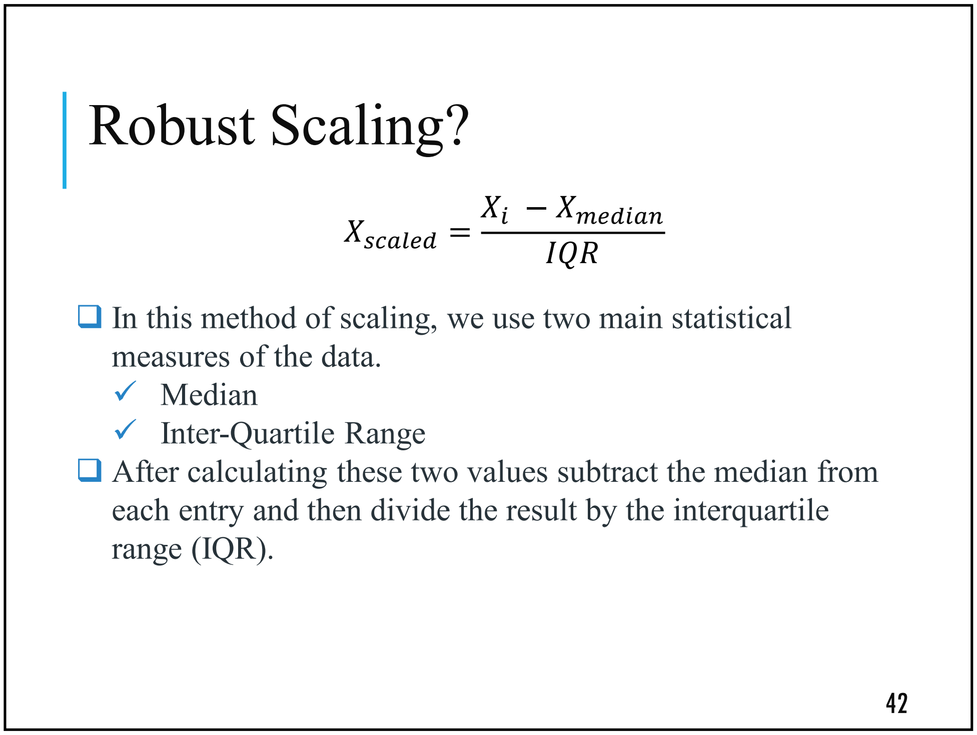
❑ This method is more or less the same as the previous method, but here instead of the minimum value, we subtract each entry by the mean value of the whole data and then divide the results by the difference between the minimum and the maximum value.



Standardization Scaling?

$$X_{scaled} = \frac{X_i - X_{mean}}{\sigma}$$

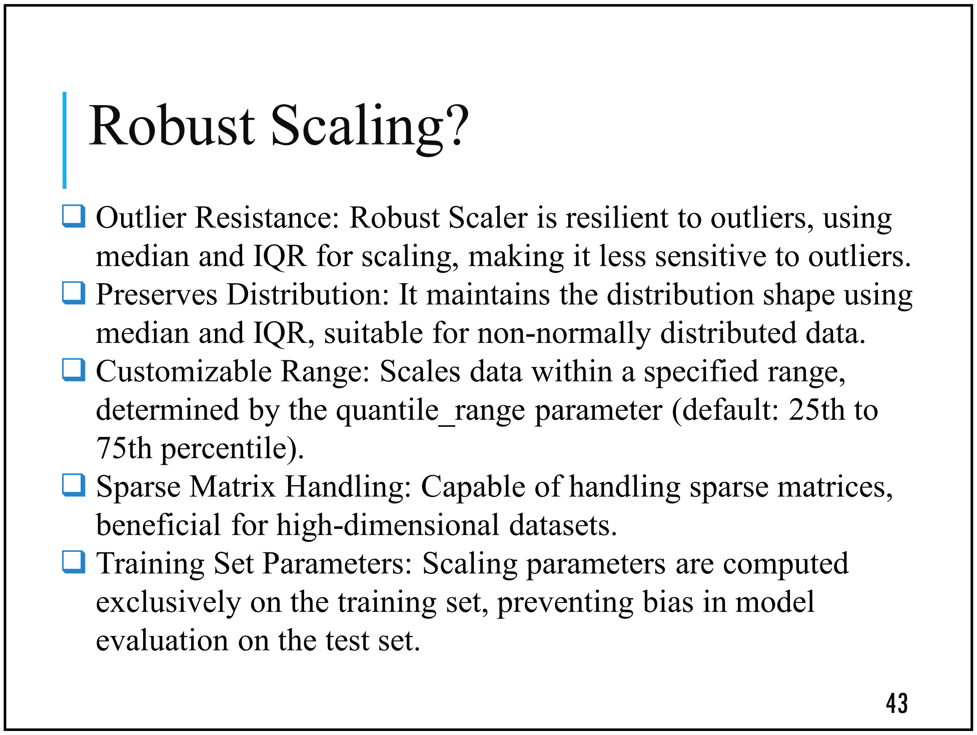
- ❑ This method of scaling is basically based on the central tendencies and variance of the data.
- ❑ First, calculate the mean and standard deviation of the data we would like to normalize.
- ❑ Then subtract the mean value from each entry and divide the result by the standard deviation.
- ❑ This helps us achieve a normal distribution (if it is already normal but skewed) of the data with a mean equal to zero and a standard deviation equal to 1.



Robust Scaling?

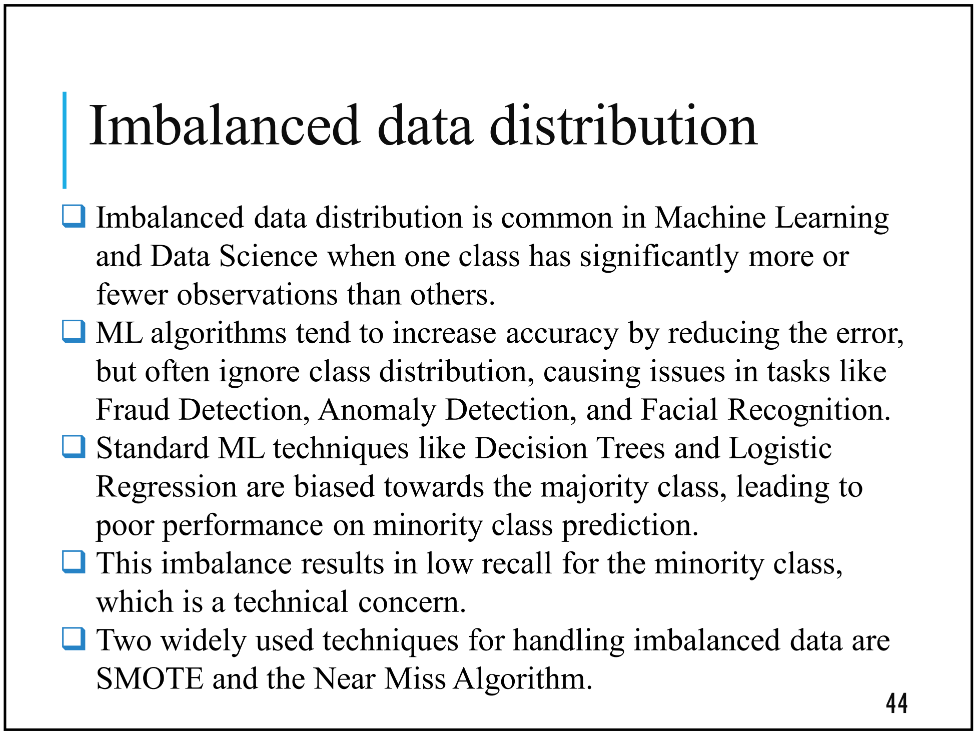
$$X_{scaled} = \frac{X_i - X_{median}}{IQR}$$

- ❑ In this method of scaling, we use two main statistical measures of the data.
  - ✓ Median
  - ✓ Inter-Quartile Range
- ❑ After calculating these two values subtract the median from each entry and then divide the result by the interquartile range (IQR).



Robust Scaling?

- Outlier Resistance: Robust Scaler is resilient to outliers, using median and IQR for scaling, making it less sensitive to outliers.
- Preserves Distribution: It maintains the distribution shape using median and IQR, suitable for non-normally distributed data.
- Customizable Range: Scales data within a specified range, determined by the quantile_range parameter (default: 25th to 75th percentile).
- Sparse Matrix Handling: Capable of handling sparse matrices, beneficial for high-dimensional datasets.
- Training Set Parameters: Scaling parameters are computed exclusively on the training set, preventing bias in model evaluation on the test set.



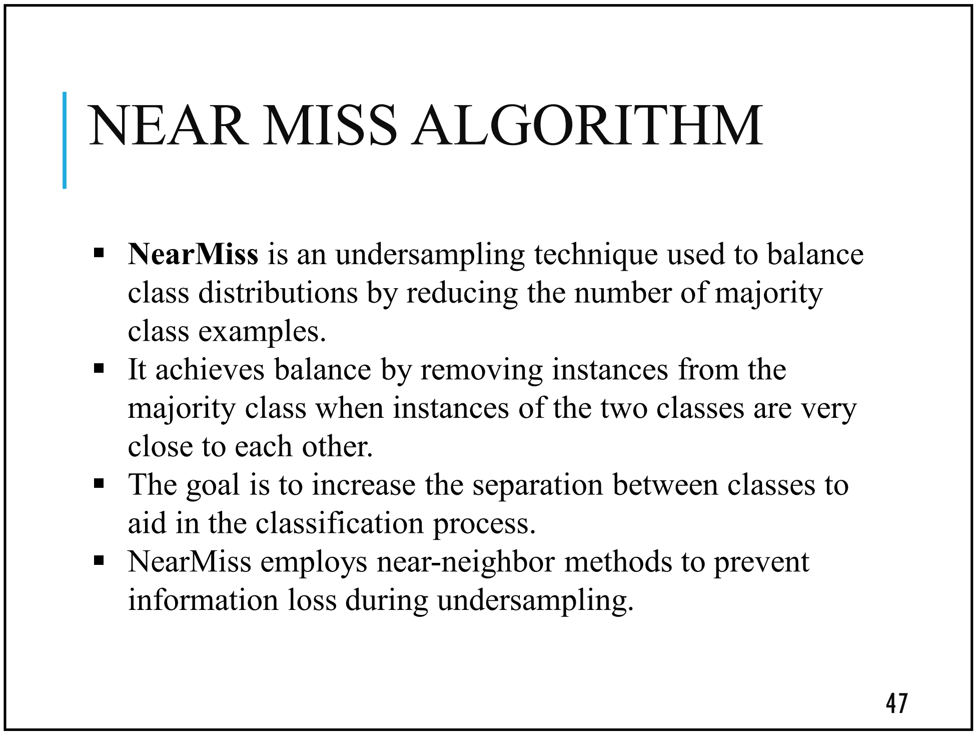
Imbalanced data distribution

- Imbalanced data distribution is common in Machine Learning and Data Science when one class has significantly more or fewer observations than others.
- ML algorithms tend to increase accuracy by reducing the error, but often ignore class distribution, causing issues in tasks like Fraud Detection, Anomaly Detection, and Facial Recognition.
- Standard ML techniques like Decision Trees and Logistic Regression are biased towards the majority class, leading to poor performance on minority class prediction.
- This imbalance results in low recall for the minority class, which is a technical concern.
- Two widely used techniques for handling imbalanced data are SMOTE and the Near Miss Algorithm.

Synthetic Minority Oversampling Technique (SMOTE)

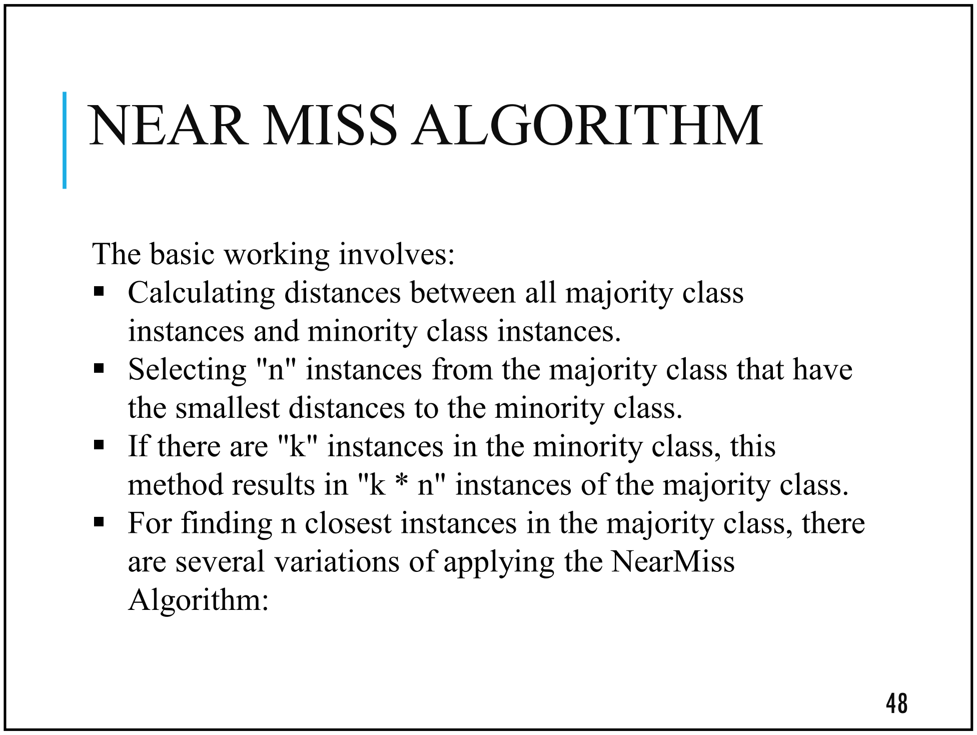
- Goal is to balance the class distribution by oversampling the minority class.
- SMOTE achieves this by creating synthetic (artificial) examples within the minority class.
- It generates these synthetic examples by interpolating between existing minority class instances.
- It selects one or more nearest neighbors for each minority class example and creates synthetic instances along the lines connecting them.
- This oversampling process helps increase the number of minority class samples, making the dataset more balanced.

Synthetic Minority Oversampling Technique (SMOTE)



NEAR MISS ALGORITHM

- **NearMiss** is an undersampling technique used to balance class distributions by reducing the number of majority class examples.
- It achieves balance by removing instances from the majority class when instances of the two classes are very close to each other.
- The goal is to increase the separation between classes to aid in the classification process.
- NearMiss employs near-neighbor methods to prevent information loss during undersampling.



NEAR MISS ALGORITHM

The basic working involves:

- Calculating distances between all majority class instances and minority class instances.
- Selecting "n" instances from the majority class that have the smallest distances to the minority class.
- If there are "k" instances in the minority class, this method results in "k * n" instances of the majority class.
- For finding n closest instances in the majority class, there are several variations of applying the NearMiss Algorithm:

NEAR MISS ALGORITHM

- NearMiss – Version 1: Selects majority class samples where the average distances to the "k" closest minority class instances are the smallest.
- NearMiss – Version 2: Selects majority class samples where the average distances to the "k" farthest minority class instances are the smallest.
- NearMiss – Version 3: Works in two steps: first, it stores the "M" nearest-neighbors for each minority class instance, and then it selects majority class instances where the average distance to the "N" nearest-neighbors is the largest.

Synthetic Minority Oversampling Technique (SMOTE)

References

https://www.geeksforgeeks.org/ml-understanding-data-processing/?ref=lbp
“Data Preparation for Data Mining” by Dorian Pyle.

QUESTIONS