Welcome to the CoGrammar Data Preprocessing

The session will start shortly...

Questions? Drop them in the chat. We'll have dedicated moderators answering questions.

CoGrammar

Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

Data Science Session Housekeeping cont.

- For all **non-academic questions**, please submit a query:
  www.hyperiondev.com/support

- Report a **safeguarding** incident:
  www.hyperiondev.com/safeguardreporting

- We would love your **feedback** on lectures: [Feedback on Lectures]()

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

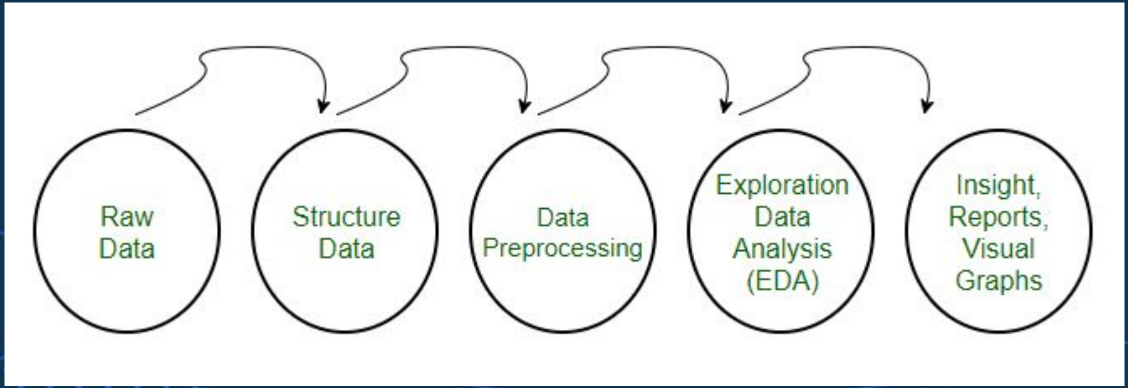
CoGrammar

Data Preprocessing

April 2024

Data Preprocessing

❖ Data preprocessing is a crucial step in the data science pipeline, going beyond basic cleaning to ensure data quality and suitability for machine learning.



Source: [GeeksForGeeks]

CoGrammar

Overview

- **Feature scaling:** Standardization, min-max scaling, robust scaling
- **Encoding categorical variables:** One-hot, label, ordinal encoding
- **Feature engineering:** Creating new features from existing data
- **Handling imbalanced data:** Oversampling, undersampling, class weights

CoGrammar

Learning objectives

❖ Understand the importance and purpose of **data preprocessing** in data science projects

❖ Learn and apply **advanced data preprocessing techniques** beyond basic data cleaning

❖ Gain hands-on experience using **Python libraries for preprocessing real-world datasets**

❖ **Integrate preprocessing techniques** into machine learning workflows

CoGrammar

Recap of Data Cleaning

CoGrammar

Data Cleaning Recap

❖ Data cleaning addresses fundamental data quality issues:

➢ **Handling missing values:** Deletion or imputation

➢ **Dealing with outliers:** Removal, transformation, or winsorization

➢ **Resolving inconsistencies:** Standardizing formats and conventions

➢ **Removing duplicates:** Eliminating redundancy

Importance of Data Preprocessing

CoGrammar

Importance

❖ **Improved data quality:** Addresses complex issues beyond basic cleaning

❖ **Enhanced model performance:** Optimizes data for learning algorithms

❖ **Reduced computational complexity:** Reduces dimensionality and creates efficient representations

CoGrammar

Feature Scaling

CoGrammar

Feature Scaling

- Purpose: Ensure fair comparison and contribution of features
- Techniques:
  - **Standardization (Z-score normalization):** Transforms features to have zero mean and unit varia

    $$X' = \frac{X - \mu}{\sigma}$$

  - **Min-max scaling:** Scales features to a specific range, typically 0 to 1

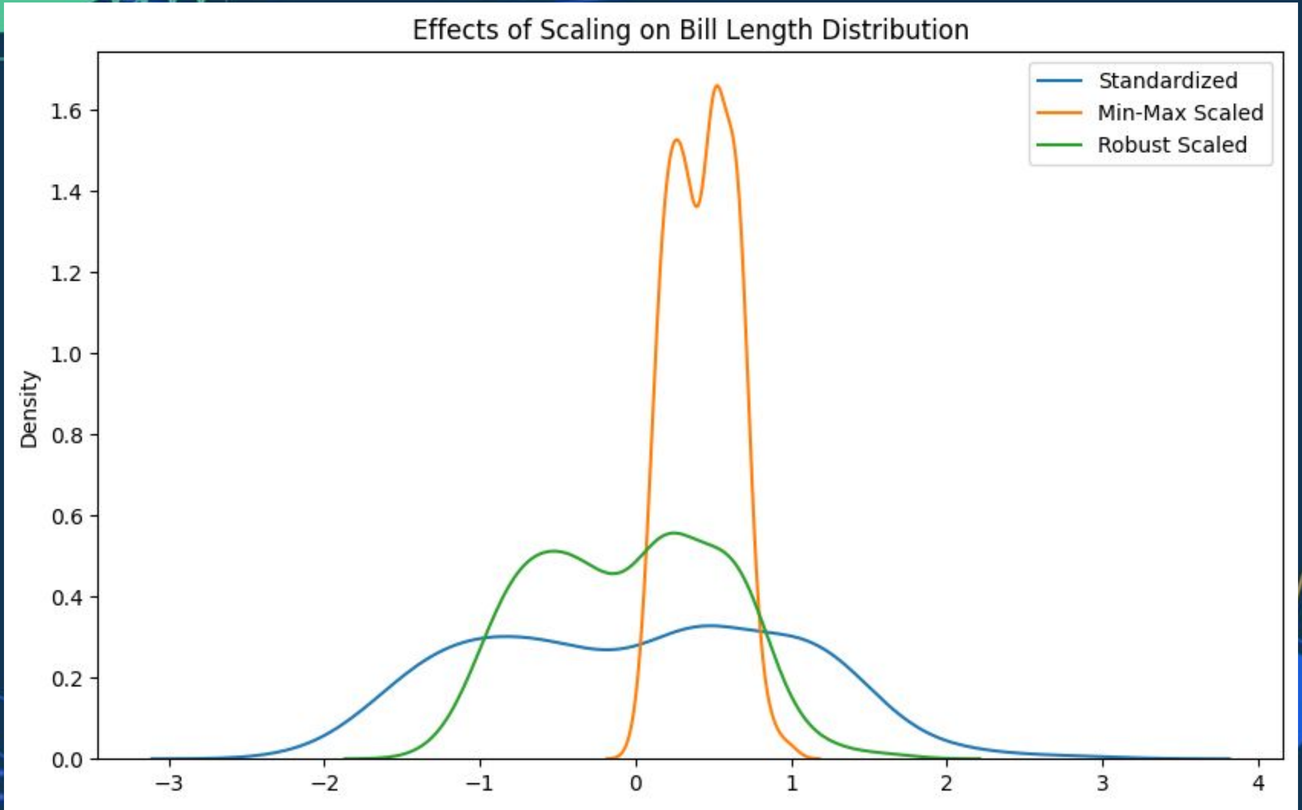
    $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  - **Robust scaling:** Uses robust statistics (median and interquartile range) to scale features
    - (X-median)/IQR

CoGrammar

Considerations

- **Standardization:** Good default, assumes Gaussian distribution
- **Min-max scaling:** Suitable for bounded features or non-Gaussian data
- **Robust scaling:** Recommended when outliers are present



Effects of Scaling on Bill Length Distribution
Standardized
Min-Max Scaled
Robust Scaled
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
4

CoGrammar

What is the purpose of feature scaling?

A. To convert categorical variables into numerical representations

B. To create new features from existing data

C. To ensure fair comparison and contribution of features in machine learning

D. To handle imbalanced class distributions

CoGrammar

What is the purpose of feature scaling?

A. To convert categorical variables into numerical representations

B. To create new features from existing data

C. To ensure fair comparison and contribution of features in machine learning

D. To handle imbalanced class distributions

CoGrammar

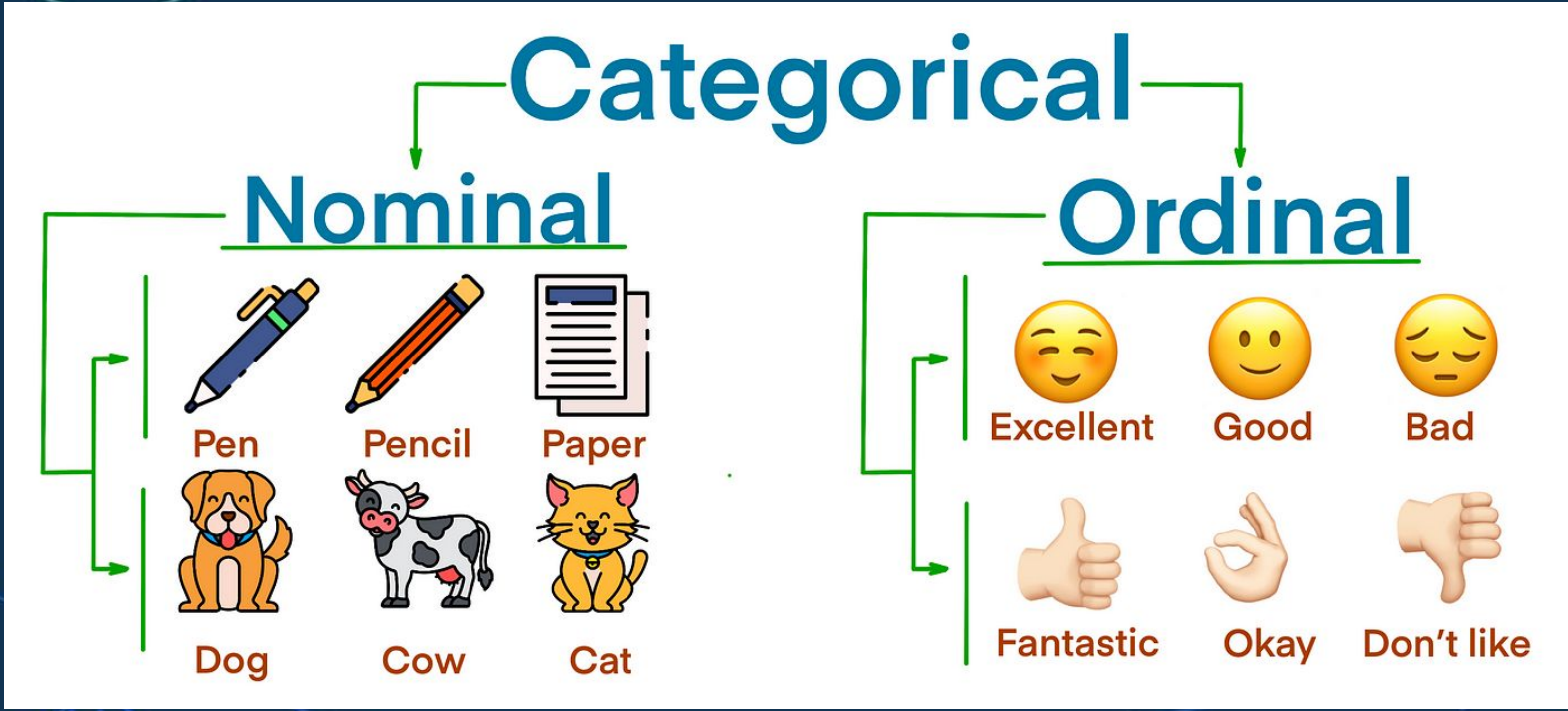
Encoding Categorical Variables

CoGrammar

Nominal vs. Ordinal

- **Nominal:** Categories without inherent order (e.g., color)
- **Ordinal:** Categories with meaningful order (e.g., size)

CoGrammar


Categorical
Nominal
Ordinal
Pen
Pencil
Paper
Dog
Cow
Cat
Excellent
Good
Bad
Fantastic
Okay
Don't like

CoGrammar

Encoding Nominal variables

- One-hot encoding: Creates binary dummy variables for each category
  - Increases dimensionality, which may impact model performance

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Source: [AnalyticsVidhya]

Encoding Nominal variables

- Binary encoding: Assigns unique binary codes to categories
  - Useful when the number of categories is large, and one-hot encoding leads to high dimensionality

	City
0	Delhi
1	Mumbai
2	Hyderabad
3	Chennai
4	Bangalore
5	Delhi
6	Hyderabad
7	Mumbai
8	Agra

	City_0	City_1	City_2	City_3
0	0	0	0	1
1	0	0	1	0
2	0	0	1	1
3	0	1	0	0
4	0	1	0	1
5	0	0	0	1
6	0	0	1	1
7	0	0	1	0
8	0	1	1	0

Source: AnalyticsVidhya

Encoding Ordinal variables

- ❖ Label encoding: Assigns numerical labels based on order
  - ➢ Maintains ordinal information but implies linear relationships between categories
  - ➢ May not be appropriate if the ordinal relationship is not linear

	Degree
0	1
1	4
2	2
3	3
4	3
5	4
6	5
7	1
8	1

Source: AnalyticsVidhya

Encoding Ordinal variables

- ❖ Ordinal encoding: Assigns numerical labels based on order
  - ➢ Preserves ordinal information without implying linear relationships
  - ➢ Suitable when the ordinal relationship between categories is meaningful

	Degree
0	1
1	4
2	2
3	3
4	3
5	4
6	5
7	1
8	1

Source: [AnalyticsVidhya]

CoGrammar

What is the main difference between nominal and ordinal variables?

A. Nominal variables have categories with an inherent order, while ordinal variables do not

B. Ordinal variables have categories with a meaningful order, while nominal variables do not

C. Nominal and ordinal variables are the same

D. Nominal variables are always encoded using one-hot encoding, while ordinal variables use label encoding

What is the main difference between nominal and ordinal variables?

A. Nominal variables have categories with an inherent order, while ordinal variables do not

B. Ordinal variables have categories with a meaningful order, while nominal variables do not

C. Nominal and ordinal variables are the same

D. Nominal variables are always encoded using one-hot encoding, while ordinal variables use label encoding

CoGrammar

Handling High-Cardinality

CoGrammar

Handling High-Cardinality

❖ The Curse of Dimensionality:

As the number of features grows, the amount of data we need to accurately be able to distinguish between these features (in order to give us a prediction) and generalize our model (learned function) grows EXPONENTIALLY.

Source: AnalyticsVidhya

Frequency-based Encoding

- Replaces categories with occurrence count
- Useful when the frequency of categories is informative

CoGrammar

Target Encoding

- Replaces categories with mean/median of target variable
- Captures the relationship between categories and the target variable

	class	Marks
0	A,	50
1	B	30
2	C	70
3	B	80
4	C	45
5	A	97
6	A	80
7	A	68

	class
0	65.000000
1	57.689414
2	59.517061
3	57.689414
4	59.517061
5	79.679951
6	79.679951
7	79.679951

Source: AnalyticsVidhya

Hashing

- Applies hash function to reduce dimensionality
- Useful when the number of categories is extremely large

	Month
0	January
1	April
2	March
3	April
4	Februay
5	June
6	July
7	June
8	September

	col_0	col_1	col_2	col_3	col_4	col_5
0	0	0	0	0	1	0
1	0	0	0	1	0	0
2	0	0	0	0	1	0
3	0	0	0	1	0	0
4	0	0	0	1	0	0
5	0	1	0	0	0	0
6	1	0	0	0	0	0
7	0	1	0	0	0	0
8	0	0	0	0	1	0

Source: AnalyticsVidhya

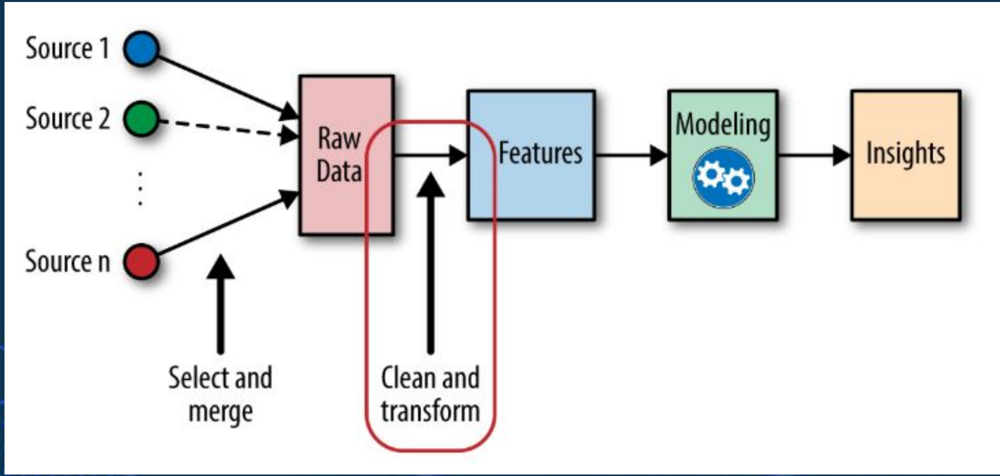
CoGrammar

Feature Engineering

CoGrammar

Feature Engineering

❖ Create informative features that improve model performance and interpretability



Source: AnalyticsVidhya

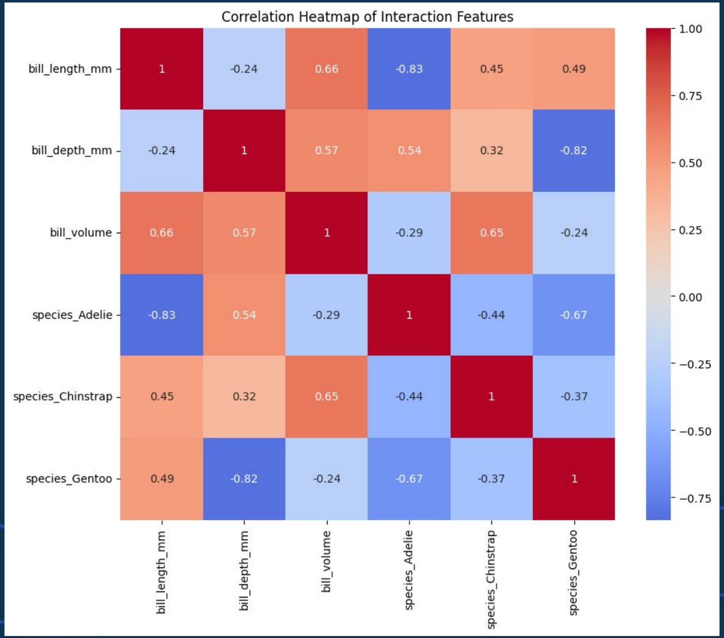
CoGrammar

Techniques

- **Interaction features:** Combine existing features to capture interactions
- **Polynomial features:** Generate higher-order terms to capture non-linear relationships
- **Domain-specific features:** Apply domain knowledge to create meaningful features

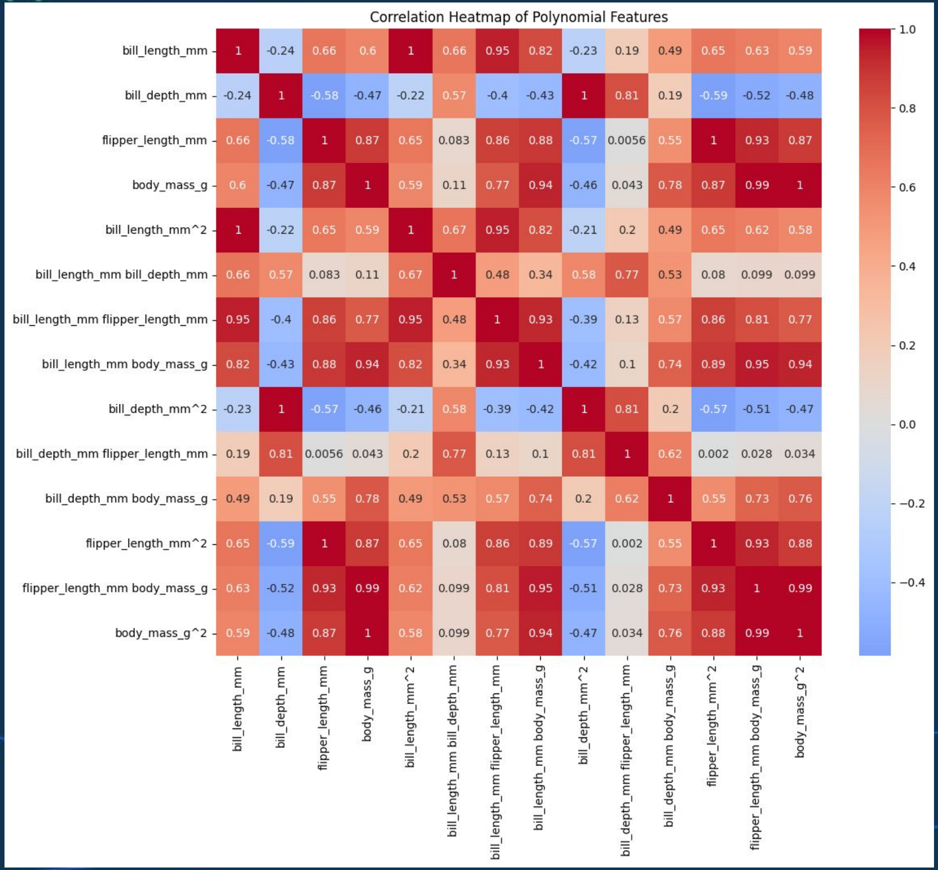
Interaction Features

```python
# Interaction features
penguins['bill_volume'] = penguins['bill_length_mm'] * penguins['bill_depth_mm']
```



CoGrammar

Polynomial Features



Handling Imbalanced Data

Challenge

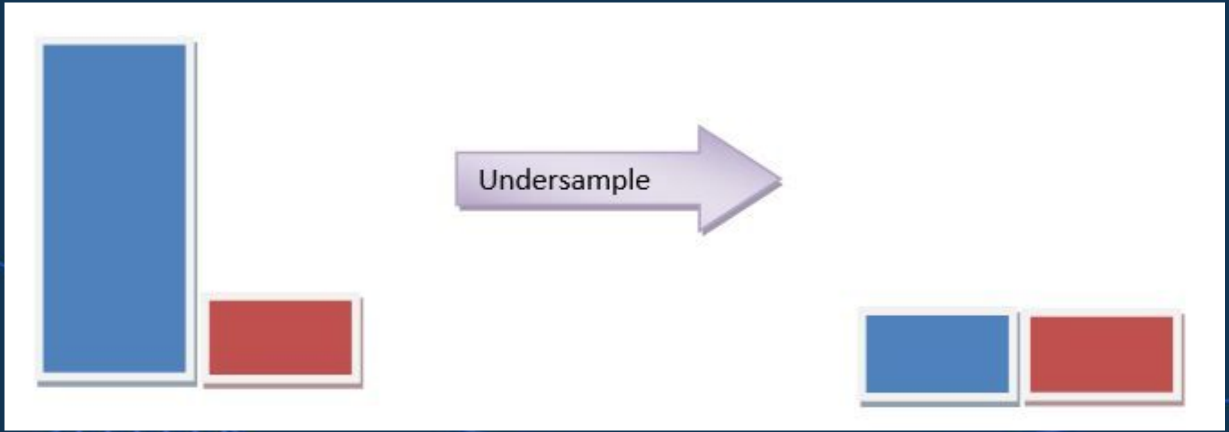
❖ Skewed class distribution leads to biased models and poor minority class performance

Source: [AnalyticsVidhya]

CoGrammar

Techniques

- **Undersampling:** Reduce majority class instances
  - **Random undersampling:** Remove majority instances



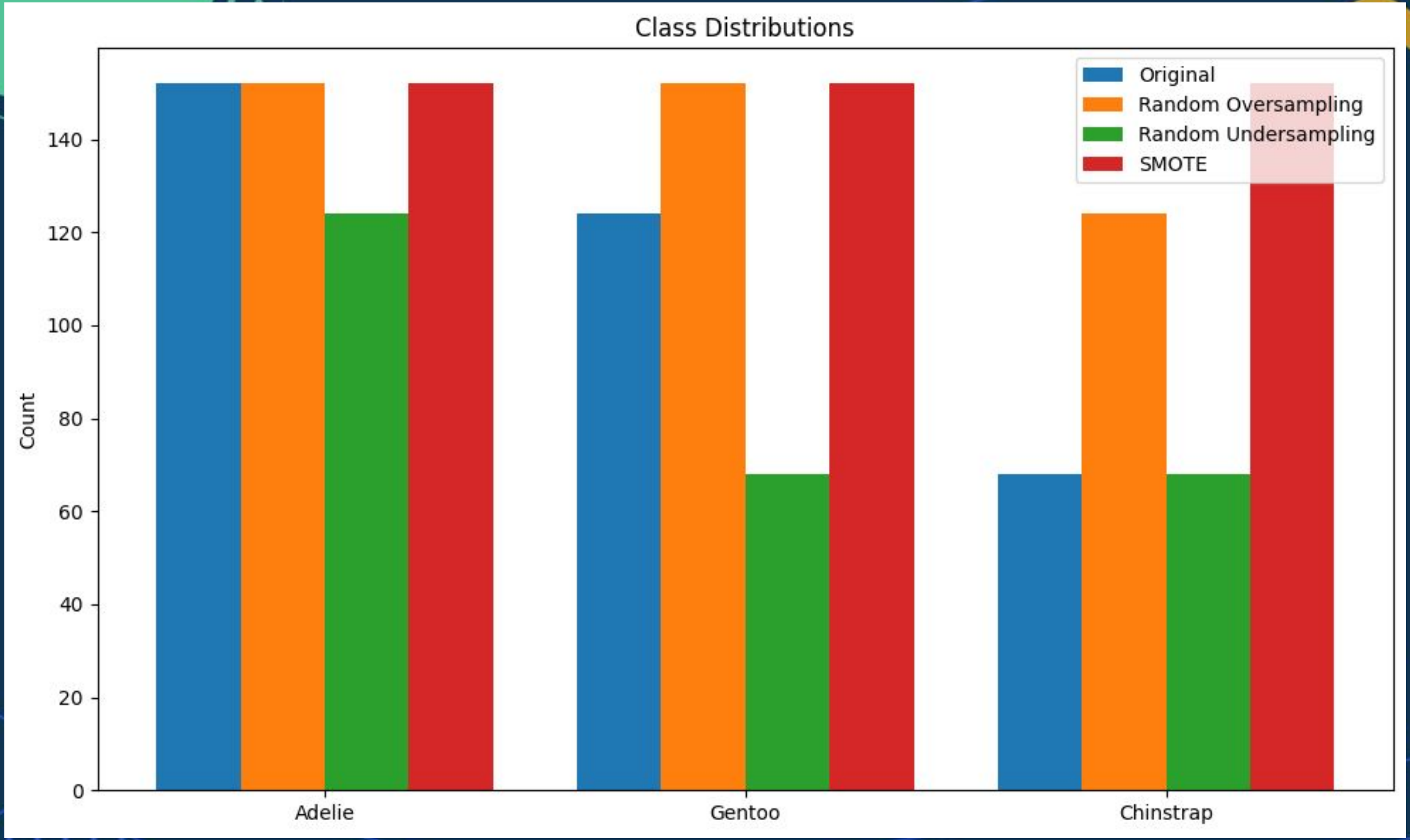
Source: [AnalyticsVidhya]

CoGrammar

Techniques

- **Oversampling:** Increase minority class instances
  - **Random oversampling:** Duplicate minority instances
  - **SMOTE:** Generate synthetic minority instances

Source: [AnalyticsVidhya]



Class Distributions
Original
Random Oversampling
Random Undersampling
SMOTE
Count
0
20
40
60
80
100
120
140
Adelie
Gentoo
Chinstrap

CoGrammar

Considerations

- Oversampling may lead to overfitting, especially with random oversampling
- Undersampling may discard potentially useful data

What is the purpose of oversampling in handling imbalanced data?

A. To reduce the number of instances in the majority class
B. To remove majority instances that are close to minority instances
C. To increase the number of instances in the minority class
D. To generate synthetic minority instances by interpolating between existing instances

What is the purpose of oversampling in handling imbalanced data?

A. To reduce the number of instances in the majority class

B. To remove majority instances that are close to minority instances

C. To increase the number of instances in the minority class

D. To generate synthetic minority instances by interpolating between existing instances

Questions and Answers

CoGrammar

Thank you for attending

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar