

Unveiling Data Insights

A. Simon Alex

BSc(CS), MSc(CS)-MUK
saatwiine@kab.ac.ug

Exploratory Data Analysis (EDA)



INDABAX
AI CLUB
Kabale University

Overview

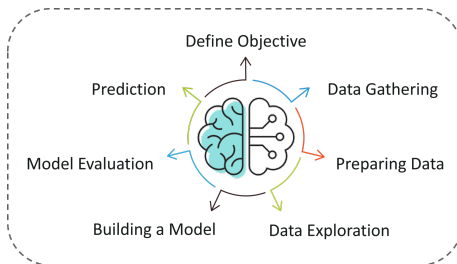
1 Unveiling Data Insights

- Machine Learning Process
- Data Gathering
- Exploring Datasets
- Handling Missing Data
- Feature Scaling & Normalization
- Encoding Categorical Data
- Feature Engineering



INDABAX
AI CLUB
Kabale University

Machine Learning Process



**INDABAX
AI CLUB**
Kabale University

Data Gathering

- Open Data Repositories:
 - Kaggle Datasets
 - Data.gov
 - World Bank Open Data
 - UCI Machine Learning



INDABAX
AI CLUB
Kabale University

Classifications of Data

- Data can be classified as **structured** or **unstructured**.
- The most common form of structured data is organized in tables with rows and columns.
- Unstructured data includes **images**, a collection of pixels, and **text**, sequences of words.



INDABAX
AI CLUB
Kabale University

Elements of Structured Data

- There are two basic types of structured data: numeric and categorical.
- Numeric data comes in two forms: **continuous**, and **discrete**.
- Categorical data takes only a fixed set of values and includes **binary** data and **ordinal** data.
- Binary data takes only two values, ordinal data on the other hand consists of categories that are ordered.



INDABAX
AI CLUB
Kabale University

Exploring Datasets

```
df = pd.read_csv('my_data.csv')  
print(df.shape)  
print(df.info())  
print(df.describe())  
print(df.head())  
print(df.tail())  
print(list(df.columns))
```



INDABAX
AI CLUB
Kabale University

Handling Missing Data

- Missing data is one of the most common problems in ML datasets.
- Missing data values can lead to **biased results** due to incomplete information and **errors in model training** as most ML algorithms require complete data.
- Missing data occurs as a result of **data entry errors**, **sensor failures**, and **user input issues**.



INDABAX
AI CLUB
Kabale University

Methods to Handle Missing Data

- Removing Rows or Columns with Missing Values
- Imputing Missing Values (Filling in Missing Data)
- Predicting Missing Values Using ML Models



INDABAX
AI CLUB
Kabale University

Handling Missing Data

```
df.isnull()
df.isnull().sum()
df.dropna()
df['Marks'].fillna(df['Marks'].mean())
df['Marks'].fillna(df['Marks'].median())
df['Marks'].fillna(df['Marks'].mode()[0])
df['Marks'].fillna(method='ffill')
df['Marks'].fillna(method='bfill')
```



INDABAX
AI CLUB
Kabale University

Feature Scaling & Normalization

- Machine learning models work best when the numerical features are on a similar scale.
- The difference in scale can make the model biased towards larger numerical values.
- The two main scaling techniques are **normalization** used when data does not have extreme outliers and **standardization** useful when data contains outliers.



INDABAX
AI CLUB
Kabale University

Encoding Categorical Data

- Machine learning models work with numerical data so categorical features must be converted into numbers.
- Methods for encoding categorical data include: label encoding, and one-hot encoding (OHE).



INDABAX
AI CLUB
Kabale University

Encoding Categorical Data

Label Encoding

Color	Label Encoded
Red	0
Blue	1
Green	2

Table: Label encoding example



INDABAX
AI CLUB
Kabale University

Encoding Categorical Data

One-hot Encoding

Color	Red	Blue	Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1

Table: One-hot encoding example



Feature Engineering

House ID	Size (sq ft)	Bedrooms	Location	Price \$
1	1,200	2	New York	250,000
2	1,500	3	San Francisco	320,000
3	800	1	Miami	180,000
4	2,000	4	Chicago	450,000

Table: House Price Dataset



INDABAX
AI CLUB
Kabale University

Feature Engineering

```
X = dataset.iloc[:, :-1].values  
y = dataset.iloc[:, -1].values
```



INDABAX
AI CLUB
Kabale University

References

- Dhinakaran, M., et al. Machine Learning For Absolute Beginners. Booksclinic Publishing, 2023.
- Bruce, Peter, Andrew Bruce, and Peter Gedeck. Practical statistics for data scientists. O'Reilly Media, 2020.



INDABAX
AI CLUB
Kabale University

Questions?



INDABAX
AI CLUB
Kabale University