# DUPLICATES HANDLING

- What is duplicate data
- Impacts on ml model
- Strategies to detect
- Handling strategies and their use cases
- Key considerations

Duplicate data is a common issue in data cleaning that can impact the accuracy and reliability of our machine learning models.

So, what is duplicate data, and why does it matter?

Duplicate data refers to multiple instances of the same observation or record in a dataset.

Different causes of duplicates are,

- **Data entry errors:** When manual data entry is involved, typos or other mistakes can result in duplicate entries.

- **System glitches:** Technical issues, such as server crashes or connection interruptions, can cause duplicate data to be created.

- **Merging of data from multiple sources:** When integrating data from multiple sources, duplicate records may be created if there is no matching process in place.

- **Lack of unique identifiers:** If there are no unique identifiers or keys in the dataset, it can be challenging to detect and remove duplicate records.

- **Incomplete data:** Incomplete data can result in duplicate records if the same information is entered in multiple fields.

- **Human error:** Even when using automated processes, mistakes can be made during data transformation, leading to duplicate records.

Impacts of duplicate data on our machine learning model are,

- **Bias:** Duplicates can bias the model towards the duplicated data, leading to incorrect predictions. For example, if there are duplicates in the training data, the model may assign more weight to those duplicates, which can skew the results.

- **Overfitting:** Duplicates can cause overfitting, where the model becomes too complex and over-adapted to the training data. This can lead to poor generalization and inaccurate predictions on new data.

- **Increased processing time:** Duplicates can increase the processing time required to train a machine learning model. This is because the model has to process more data, which can slow down the training process and make it more computationally expensive.

- **Reduced data quality:** Duplicates can reduce the quality of the dataset used to train the model, which can lead to incorrect or unreliable predictions.

- **Data imbalance:** Duplicates can create data imbalance, where certain data points are overrepresented in the dataset, leading to an imbalanced model. This can lead to incorrect predictions, particularly for minority classes.

- **Redundancy:** Duplicates can add redundancy to the dataset, which can increase the complexity of the model unnecessarily. This can lead to longer training times and potentially worse performance.

- **Inconsistency:** Duplicates can introduce inconsistencies in the data, particularly if the duplicates have different values for the same features. This can lead to confusion for the model and result in inaccurate predictions.

Now, How can we detect duplicates?
Here are some popular strategies to detect duplicates,

**Exact match**

- This strategy involves comparing each record in the dataset to every other record to find exact duplicates.
- This is a simple and effective approach that works well for small datasets.

**Fuzzy matching**

- Fuzzy matching is a technique used to identify duplicates that are not exact matches but are very similar.
- This can be done using algorithms that compare the similarity of strings or values.
- Fuzzy matching is useful when dealing with messy data, such as misspelled or abbreviated values.

## Hashing

- Hashing is a technique that converts data into a unique string of characters that can be used to identify duplicates.
- It is particularly useful for very large datasets, as it allows for efficient comparison of records without having to compare each record to every other record.

## Machine learning techniques

- Machine learning techniques such as clustering can be used to identify groups of records that are similar and may be potential duplicates.
- This approach is useful when dealing with large datasets where manual inspection may be impractical.

The specific strategy chosen will depend on the characteristics of the dataset and the desired outcome of the data cleaning process.

So, What are the different strategies used to handle the duplicate data?
Here are the some more popular strategies,

**Removing duplicates**

- The most straightforward strategy is to remove the duplicate records from the dataset.
- This approach can be effective if the duplicates are not providing any additional information.

**Aggregating duplicates**

- If the duplicates contain slightly different information or values, it may be appropriate to aggregate them.
- For example, if multiple records contain different purchase amounts for the same customer, the records could be aggregated by calculating the total purchase amount.

**Merging duplicates**

- When dealing with multiple datasets, it may be necessary to merge duplicate records from different sources.
- This approach requires careful matching and deduplication to ensure that the correct records are merged.

**Keeping only one copy of the duplicate record**

- Another approach is to keep only one copy of the duplicate record and delete the others.
- This can be done by selecting the most recent or the highest quality record.
-

**Investigating the cause of duplicates**

- It's also important to investigate the cause of duplicates to prevent them from occurring in the future.
- This may involve implementing data validation checks or improving data entry processes.

By using these strategies, we can effectively handle duplicates in our datasets, improving the accuracy and reliability of our machine learning models.

And one more thing,
While handling duplicates in data cleaning, there are several key considerations to keep in mind. These include:

- **Identify the root cause:** It's important to identify the root cause of duplicates to prevent them from occurring in the future. The duplicates may be caused by data entry errors, data merging issues, or other reasons.

- **Determine the impact on the analysis:** Duplicates can have a significant impact on data analysis, as they can skew the results and lead to inaccurate conclusions. Before removing duplicates, it's important to consider the impact they may have on the analysis and whether removing them will affect the accuracy of the results.

- **Use appropriate data cleaning techniques:** It's important to choose the appropriate technique based on the specific needs of the project and the characteristics of the data.

- **Document the decisions made:** When handling duplicates, it's important to document the decisions made and the reasoning behind them. This will help ensure that the data cleaning process is transparent and replicable.

By keeping these considerations in mind, we can effectively handle duplicates in datasets and ensure that our machine learning models are accurate and reliable.

In summary, duplicates in data used for machine learning can negatively impact the accuracy, reliability, and efficiency of the model.

It's important to identify and remove duplicates during data cleaning to ensure the best possible results from the machine learning model.

Finally, Don't let duplicates ruin your models – start detecting and handling them today

Happy learning!😊

# THANK YOU

@ Gangadhar Neelam