

DATAWARS TASK 2:

DATAWARS

Feature engineering / Dealing with duplicates

Dataset ▾

1

Check if any of the record of the dataset is duplicated

There are 5 duplicated rows

There are not duplicated rows

There are 9 duplicated rows

There are 29 duplicated rows

Submitted

Correct!

Report issue

Checking for duplicates in a specific column is an essential step in data preprocessing and analysis. It ensures data accuracy, reliability, and consistency, which are fundamental for drawing meaningful conclusions and making informed decisions based on the data.

2

Your task is to check for duplicates in a specific column of a given DataFrame. Select the correct code

df.duplicated('TypeText',axis=1)

df.duplicated('TypeText')

df.duplicated('TypeText')

Submitted

Correct!

Report issue

3

Drop duplicated

Your task is to drop duplicate rows from a given DataFrame and retain the first occurrence of each duplicated row. Select the correct code.

df.drop_duplicates(keep='last',inplace=True)

df.drop_duplicates('first')

df.drop_duplicates(keep='first',inplace=True)

df.dropna()

Submitted

Correct!

Report issue

Quiz

Completed activities 5/5 100%

Quiz ▾

4

Scenario question

You work as a data analyst for an e-commerce company that sells a wide range of products. The company has provided you with a dataset of customer reviews, which includes attributes such as the product purchased, review text, rating, and demographic information of customers. Your task is to leverage this data to enhance the product recommendation system.

Scenario: You have conducted an initial analysis of the customer review dataset. During this analysis, you identified that there are several duplicate entries in the dataset, which can lead to biased results and flawed conclusions. What should be your next step to address this issue?

Implement data preprocessing techniques to identify and remove duplicate entries effectively.

Merge the duplicate entries into single records to increase the dataset's size for more accurate recommendations.

Leave the duplicates as is and conduct a separate analysis on them to explore potential patterns.

Submitted

Correct!

Report issue

Practice validation curve ▾

Quiz

1

True or False: Overfitting occurs when a model is too complex and fits the training data too closely, including noise. This can lead to poor generalization on new, unseen data.

False

True

Submitted

Correct!

Report issue

2

True or False: Validation curves help in determining the best hyperparameters for a model by plotting the training and testing error against different values of the hyperparameter.

False

True

Submitted

Correct!

Report issue

Characteristics of Good Features in Machine Learning ▾

Quiz

Here are four true or false questions related to the importance of feature selection and engineering in machine learning. Let's do it!

- 1 True or False: Feature selection is the process of randomly choosing any subset of features from the dataset.

☐ True
☒ False

Submitted Correct!

[Report issue](#)

- 2 True or False: Dimensionality reduction is a technique used in feature engineering to increase the number of features in the dataset.

☐ True
☒ False

Submitted Correct!

[Report issue](#)

Quiz ▾

[Report issue](#)

- 3 True or False: Feature engineering involves creating new features by combining or transforming existing ones to enhance the model's performance.

☐ False
☒ True

Submitted Correct!

[Report issue](#)

- 4 True or False: Features that exhibit high discriminative power are typically highly correlated with each other.

☒ False
☐ True

Submitted Correct!

[Report issue](#)

Practice ▾

- 1 Which of the following is NOT a machine learning model used in the exercise?

☒ Linear Regression
☐ Support Vector Machine (SVM)
☐ Neural Network
☐ Random Forest

Submitted Correct!

[Report issue](#)

- 2 How can changing hyperparameter values impact a model's performance?

☒ It can lead to both improvements and deteriorations in performance.
☐ It has no effect on the model's performance.
☐ It only affects the model's visualization.
☐ It can only improve the model's performance.

Submitted Correct!

[Report issue](#)

Practice ▾

[Report issue](#)

- 3 True or False: The accuracy of the SVM with linear kernel model decreases with low values of C.

☐ False
☒ True

Submitted Correct!

[Report issue](#)

- 4 True or False: The accuracy of the Random Forest model decreases with low values of max_depth (with max_depth=2 and 5).

☒ False
☐ True

Submitted Correct!

[Report issue](#)

