Certified Professional in AI Applications in Aviation

Data Preprocessing and Feature Engineering

Data Preprocessing: The process of cleaning, transforming, and preparing raw data for analysis and modeling. It includes various techniques such as data cleaning, data transformation, data normalization, and data reduction. The goal is to create a high-quality dataset that can improve the performance of machine learning algorithms.

Feature Engineering: The process of creating new features or transforming existing features to improve the performance of machine learning algorithms. It includes various techniques such as feature scaling, feature extraction, feature selection, and dimensionality reduction. The goal is to create a feature set that can capture meaningful patterns and relationships in the data.

Data Cleaning: The process of identifying and correcting errors, inconsistencies, and missing values in the data. It includes various techniques such as data imputation, outlier detection, and noise reduction. The goal is to create a clean dataset that can improve the accuracy and reliability of machine learning algorithms.

Data Transformation: The process of converting data from one format to another to make it suitable for analysis and modeling. It includes various techniques such as data encoding, data aggregation, and data discretization. The goal is to create a transformed dataset that can improve the interpretability and visualization of machine learning algorithms.

Data Normalization: The process of scaling numeric data to a common range to improve the performance of machine learning algorithms. It includes various techniques such as min-max scaling, z-score scaling, and decimal scaling. The goal is to create a normalized dataset that can avoid bias and improve the convergence of machine learning algorithms.

Data Reduction: The process of reducing the dimensionality of the data to improve the efficiency and effectiveness of machine learning algorithms. It includes various techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), and factor analysis. The goal is to create a reduced dataset that can capture the essential features of the data and avoid overfitting.

Feature Scaling: The process of transforming the scale of the features to a common range to improve the performance of machine learning algorithms. It includes various techniques such as min-max scaling, z-score scaling, and decimal scaling. The goal is to create a scaled feature set that can avoid bias and improve the convergence of machine learning algorithms.

Feature Extraction: The process of creating new features by combining or transforming existing features to improve the performance of machine learning algorithms. It includes various techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA). The goal is to create a feature set that can capture meaningful patterns and relationships in the data.

Feature Selection: The process of selecting a subset of relevant features from the original feature set to improve the performance of machine learning algorithms. It includes various techniques such as filter methods, wrapper methods, and embedded methods. The goal is to create a selected feature set that can avoid overfitting and improve the interpretability of machine learning algorithms.

Dimensionality Reduction: The process of reducing the number of features in the dataset to improve the efficiency and effectiveness of machine learning algorithms. It includes various techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), and singular value decomposition (SVD). The goal is to create a reduced dataset that can capture the essential features of the data and avoid overfitting.

Data Imputation: The process of replacing missing values with estimated values to create a complete dataset for analysis and modeling. It includes various techniques such as mean imputation, median imputation, mode imputation, and regression imputation. The goal is to create an imputed dataset that can avoid bias and improve the accuracy of machine learning algorithms.

Outlier Detection: The process of identifying and handling data points that are significantly different from other data points in the dataset. It includes various techniques such as statistical methods, distance-based methods, and density-based methods. The goal is to create an outlier-free dataset that can avoid bias and improve the accuracy of machine learning algorithms.

Noise Reduction: The process of removing or reducing the impact of noise in the data to create a clean dataset for analysis and modeling. It includes various techniques such as smoothing, filtering, and wavelet transform. The goal is to create a noise-free dataset that can improve the accuracy and reliability of machine learning algorithms.

Data Encoding: The process of converting categorical data into numerical data to make it suitable for analysis and modeling. It includes various techniques such as label encoding, one-hot encoding, and binary encoding. The goal is to create an encoded dataset that can

improve the performance of machine learning algorithms.

Data Aggregation: The process of combining multiple data points into a single data point to create a summarized dataset for analysis and modeling. It includes various techniques such as mean aggregation, median aggregation, and mode aggregation. The goal is to create an aggregated dataset that can improve the interpretability and visualization of machine learning algorithms.

Data Discretization: The process of converting continuous data into discrete data to make it suitable for analysis and modeling. It includes various techniques such as equal width discretization, equal frequency discretization, and clustering-based discretization. The goal is to create a discretized dataset that can improve the interpretability and visualization of machine learning algorithms.

Min-Max Scaling: A feature scaling technique that scales the data to a common range between 0 and 1. It is calculated as (x - min(x)) / (max(x) - min(x)). The goal is to create a scaled dataset that can avoid bias and improve the convergence of machine learning algorithms.

Z-Score Scaling: A feature scaling technique that scales the data to a common range with a mean of 0 and a standard deviation of 1. It is calculated as (x - mean(x)) / std(x). The goal is to create a scaled dataset that can avoid bias and improve the convergence of machine learning algorithms.

Decimal Scaling: A feature scaling technique that scales the data by dividing each value by a power of 10. It is calculated as x / 10^n, where n is the smallest integer that makes the maximum absolute value less than 1. The goal is to create a scaled dataset that can avoid bias and improve the convergence of machine learning algorithms.

Principal Component Analysis (PCA): A dimensionality reduction technique that transforms the original features into a new set of features called principal components. It captures the most important patterns and relationships in the data while reducing the number of features. The goal is to create a reduced dataset that can improve the efficiency and effectiveness of machine learning algorithms.

Linear Discriminant Analysis (LDA): A feature extraction technique that transforms the original features into a new set of features called discriminant functions. It captures the most important differences between classes while reducing the number of features. The goal is to create a feature set that can improve the classification accuracy of machine learning algorithms.

Filter Methods: A feature selection technique that evaluates each feature independently based on a statistical or information-theoretic criterion. It includes various measures such as correlation, mutual information, and chi-square test. The goal is to create a selected feature set that can avoid overfitting and improve the interpretability of machine learning algorithms.

Wrapper Methods: A feature selection technique that evaluates each feature subset as a whole based on the performance of a specific machine learning algorithm. It includes various search strategies such as forward selection, backward elimination, and recursive feature elimination. The goal is to create a selected feature set that can optimize the performance of machine learning algorithms.

Embedded Methods: A feature selection technique that integrates the feature selection process into the machine learning algorithm itself. It includes various approaches such as regularization, sparsity, and ensemble methods. The goal is to create a selected feature set that can avoid overfitting and improve the interpretability of machine learning algorithms.

Label Encoding: A data encoding technique that assigns a unique integer value to each category in the categorical data. It includes various methods such as ordinal encoding and one-hot encoding. The goal is to create an encoded dataset that can improve the performance of machine learning algorithms.

One-Hot Encoding: A data encoding technique that creates a binary vector for each category in the categorical data. It includes various methods such as binary encoding and hashing encoding. The goal is to create an encoded dataset that can improve the performance of machine learning algorithms.

Binary Encoding: A data encoding technique that represents each category in the categorical data as a binary number. It includes various methods such as base-2 encoding and base-1