Define/Explain

Please download this document and upload to the drive when completed. This should be done individually. You can use any resources.

**Supervised/Unsupervised learning**

Supervised learning is a type of machine learning where the algorithm learns from labeled training data. The input data includes both the features (independent variables) and their corresponding labels (dependent variable or target variable). The algorithm learns to predict the labels for unseen data based on the patterns and relationships observed in the labeled training data.

Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data. In unsupervised learning, there are no predefined labels or target variables. The algorithm explores the patterns, structures, and relationships within the data to uncover hidden insights, group similar data points, or identify anomalies without prior knowledge of the expected outcomes.

**Overfitting/Underfitting**

Overfitting occurs when a machine learning model learns the training data too well, capturing noise or random fluctuations in the data instead of the underlying patterns. An overfitted model may perform very well on the training data but performs poorly on unseen data because it has not generalized well. Overfitting can be caused by having too complex a model or insufficient data for training.

Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in the training data. An underfitted model performs poorly both on the training data and unseen data because it fails to capture the important relationships in the data. Underfitting can be caused by using a too simple model or not having enough features to represent the data adequately.

**Regression**

Regression is a type of supervised learning used to predict continuous numeric values (dependent variables) based on input features (independent variables). It models the relationship between the independent variables and the dependent variable, allowing predictions and understanding of how changes in the independent variables impact the dependent variable.

**Categorical**

Categorical data refers to variables that represent discrete and distinct categories or groups. These variables can be qualitative or nominal in nature, such as colors, gender, or types of products. Categorical variables are typically represented using labels or codes rather than numerical values.

**Binary**

Binary refers to a categorical variable that has only two possible outcomes or categories. For example, it can represent yes/no, true/false, or 0/1 choices.

**Error/Loss**

Error or loss refers to the discrepancy between the predicted values of the model and the actual values in the training or validation data. It quantifies how well the model is performing by measuring the difference between predicted and actual values. The goal is to minimize the error or loss function to improve the model's accuracy.

**Normalization**

Normalization is a preprocessing technique used to standardize or rescale the features in a dataset to a common scale. It ensures that all features have similar magnitudes and reduces the impact of differences in the measurement units or scales. Common normalization methods include min-max scaling and z-score standardization.

**Dummy Columns**

Dummy columns (also called one-hot encoding) are a representation of categorical variables as binary vectors. Each unique category or label in the categorical variable is represented by a separate binary column. The presence of a category is indicated by 1, while the absence is indicated by 0. This encoding is useful for including categorical variables in machine learning models that require numerical input.

*Please add 2 more terms that we learned about:*

**Intercept**

The intercept is the value of the dependent variable when all independent variables are zero. It represents the baseline or starting point of the dependent variable. The intercept is also known as the constant term or the y-intercept in the equation of a straight line (y = mx + b).

**Coefficients**

Coefficients represent the weights or slopes assigned to the independent variables. They quantify the impact or contribution of each independent variable on the dependent variable. Each independent variable has its own coefficient, indicating the change in the dependent variable for a unit change in that particular independent variable, assuming all other variables are held constant. Coefficients can be positive or negative, indicating a positive or negative relationship between the independent variable and the dependent variable.