**Validation Set:**

Validation set là một phần của dữ liệu được sử dụng trong quá trình huấn luyện mô hình machine learning. Khi chia dữ liệu thành các tập huấn luyện và tập kiểm tra, ta thường dùng tập huấn luyện để mô hình học từ dữ liệu và sử dụng tập kiểm tra để đánh giá hiệu suất. Tuy nhiên, để điều chỉnh các siêu tham số của mô hình một cách tốt nhất, chúng ta cần một tập dữ liệu khác gọi là validation set. Tập này được sử dụng để đánh giá hiệu suất của mô hình trong quá trình huấn luyện, từ đó giúp điều chỉnh và lựa chọn siêu tham số một cách tốt nhất.

**Imbalanced Data:**

Dữ liệu mất cân đối (imbalanced data) xuất hiện khi số lượng các mẫu trong các lớp khác nhau của biến mục tiêu trong bộ dữ liệu là không đồng đều. Điều này có thể xảy ra trong các bài toán phân loại khi một lớp chiếm đa số trong khi lớp còn lại chỉ chiếm một phần nhỏ. Khi đối mặt với dữ liệu mất cân đối, mô hình có thể học không tốt lớp thiểu số vì nó có xu hướng chọn lớp đa số. Để xử lý vấn đề này, có thể sử dụng các kỹ thuật như oversampling (tăng cường mẫu trong lớp thiểu số), undersampling (giảm mẫu trong lớp đa số), hoặc sử dụng các thuật toán hoặc phương pháp đặc biệt được thiết kế để xử lý dữ liệu mất cân đối.

**Cross-Validation:**

Cross-validation là một phương pháp đánh giá hiệu suất của mô hình trong machine learning. Phương pháp này chia dữ liệu thành các tập con, sau đó huấn luyện mô hình trên các tập con này và đánh giá hiệu suất trên tập còn lại. Kỹ thuật phổ biến nhất là k-fold cross-validation, trong đó dữ liệu được chia thành k phần, mô hình được huấn luyện trên k-1 phần và đánh giá trên phần còn lại. Quá trình này được lặp lại k lần để đảm bảo mô hình được đánh giá trên toàn bộ dữ liệu. Cross-validation giúp ước lượng hiệu suất của mô hình một cách khách quan hơn khi chỉ sử dụng một tập kiểm tra duy nhất.