**ASSIGNMENT-1:**

**1. Balanced Dataset**

A **balanced dataset** is one where the different classes or categories are represented approximately equally. Each class has roughly the same number of instances.

**Example of a Balanced Dataset:**

Imagine you have a dataset for a binary classification problem where the goal is to predict if a person will buy a product (yes/no). If the dataset has 500 people who bought the product (class "yes") and 500 people who didn't buy the product (class "no"), this is a balanced dataset.

| **Buy Product (Label)** | **Count** |
| --- | --- |
| Yes | 500 |
| No | 500 |

**2. Unbalanced Dataset**

An **unbalanced dataset** (also called imbalanced dataset) is one where the classes or categories are not equally represented. One class has significantly more instances than the other(s). This can cause problems because machine learning algorithms might be biased toward the majority class and perform poorly for the minority class.

**Example of an Unbalanced Dataset:**

Now, consider the same problem, but this time you have 950 people who didn’t buy the product (class "no") and only 50 people who bought the product (class "yes"). This would be an unbalanced dataset.

| 1. **Buy Product (Label)** | 1. **Count** |
| --- | --- |
| 1. Yes | 1. 50 |
| 1. No | 1. 950 |

In this case, a model might predict "No" all the time because it’s the majority class, which might make it seem like the model is performing well, even though it isn't identifying the "Yes" class effectively.

**Why it matters:**

* **Balanced datasets** are easier to work with because algorithms can learn equally about both classes, leading to a more fair and accurate model.
* **Unbalanced datasets** require special techniques, like oversampling the minority class (e.g., SMOTE), undersampling the majority class, or adjusting the decision threshold to improve performance for the minority class.
* **ASSIGNMENT-2:**

Business understanding in data analytics refers to the process of gaining a deep understanding of the business objectives, challenges, and needs before diving into data analysis. It helps ensure that the data analytics work is aligned with the business goals, so the insights generated will be relevant and actionable.

**Examples of Business Understanding in Data Analytics:**

1. **Sales Increase for a Retail Company:**
   * **Business Goal:** The company wants to increase sales.
   * **Analytics Question:** What factors influence customer purchasing behavior?
   * **Understanding:** Knowing the goal is to increase sales, the data team might look at trends like seasonality, product preferences, or customer demographics.
2. **Customer Retention for a Telecom Company:**
   * **Business Goal:** The company wants to reduce churn (customers leaving).
   * **Analytics Question:** What factors contribute to customers leaving?
   * **Understanding:** The data team would focus on variables like service quality, pricing, and customer support interactions to predict and reduce churn.
3. **Marketing Campaign Effectiveness:**
   * **Business Goal:** A company wants to measure the success of a marketing campaign.
   * **Analytics Question:** Did the campaign lead to an increase in new customers or sales?
   * **Understanding:** Before analysis, the team needs to know what metrics (such as new sign-ups, conversions, or social media engagement) will define "success."
4. **Product Launch Decision for a Software Company:**
   * **Business Goal:** A company is deciding whether to launch a new product.
   * **Analytics Question:** What does the market need and how well can we meet that need?
   * **Understanding:** The team might explore customer feedback, market demand, and competitor performance to inform whether the product should be launched or tweaked.
5. **ASSIGNMENT-3:**

A **confusion matrix** is a table used to evaluate the performance of a classification model in data analytics. It compares the actual values (true labels) with the predicted values from the model. This helps to identify how well the model is performing and where it might be making mistakes.

**Structure of a Confusion Matrix:**

A confusion matrix typically has 4 key elements:

|  | **Predicted Positive** | **Predicted Negative** |
| --- | --- | --- |
| **Actual Positive** | True Positive (TP) | False Negative (FN) |
| **Actual Negative** | False Positive (FP) | True Negative (TN) |

* **True Positive (TP):** The model correctly predicted a positive outcome.
* **True Negative (TN):** The model correctly predicted a negative outcome.
* **False Positive (FP):** The model incorrectly predicted a positive outcome (a "false alarm").
* **False Negative (FN):** The model incorrectly predicted a negative outcome (missed a positive).

**Examples in Data Analytics:**

1. **Email Spam Filter:**
   * **Business Goal:** Identify whether an email is spam or not.
   * **Confusion Matrix:**
     + **True Positive (TP):** The model correctly identifies a spam email as spam.
     + **True Negative (TN):** The model correctly identifies a non-spam email as non-spam.
     + **False Positive (FP):** The model incorrectly marks a non-spam email as spam.
     + **False Negative (FN):** The model misses a spam email and marks it as non-spam.
2. **Medical Diagnosis (e.g., Detecting Cancer):**
   * **Business Goal:** Predict whether a patient has cancer or not.
   * **Confusion Matrix:**
     + **True Positive (TP):** The model correctly predicts a cancer diagnosis.
     + **True Negative (TN):** The model correctly predicts no cancer.
     + **False Positive (FP):** The model incorrectly diagnoses cancer when the patient doesn’t have it.
     + **False Negative (FN):** The model misses detecting cancer when the patient actually has it.
3. **Customer Churn Prediction:**
   * **Business Goal:** Predict if a customer will leave (churn) or stay.
   * **Confusion Matrix:**
     + **True Positive (TP):** The model correctly predicts a customer will churn.
     + **True Negative (TN):** The model correctly predicts a customer will stay.
     + **False Positive (FP):** The model incorrectly predicts a customer will churn when they actually stay.
     + **False Negative (FN):** The model misses a customer who will churn.

**Why is the Confusion Matrix Useful?**

It helps you understand not just the overall accuracy of the model but also where the model might be making specific types of errors (like false positives or false negatives). For example, in medical diagnosis, missing a positive case (false negative) can be far more serious than a false alarm (false positive).