ASSIGNMENT-1

1. Balanced Dataset: -

*A dataset is **balanced** when the number of samples in each class is approximately equal.

*A **balanced dataset** is a dataset where all classes in a classification problem have an approximately equal number of samples.

Example:

- A dataset with **500 positive (Yes)** and **500 negative (No)** samples in a binary classification problem.
- A multi-class dataset with three classes, each having around **1,000 samples**.

Advantages:

- No inherent bias towards any class.
- Machine learning models learn better patterns.
- Improves the accuracy of predictions across all classes.

Use Cases:

- Spam vs. Non-Spam classification.
- Sentiment Analysis (Positive, Negative, Neutral) with equal representation.
- Image classification where all categories have similar sample counts.

Main example:-

Example: Balanced Dataset for Disease Classification

Patient ID	Symptoms	Diagnosis
101	Fever, Cough, Fatigue	Flu
102	Chest Pain, Shortness of Breath	Heart Disease
103	Skin Rash, Itching	Allergy
104	Fever, Chills, Body Aches	Flu
105	Chest Discomfort, Dizziness	Heart Disease

106	Sneezing, Runny Nose	Allergy

2. Unbalanced Dataset

*A dataset is **unbalanced** when one class has significantly more samples than others.

*An **unbalanced dataset** (also called **imbalanced dataset**) is a dataset where one class has significantly more samples than the other(s). This occurs mainly in **classification problems** and can cause machine learning models to be biased towards the majority class.

⍰ **Example:**

- Fraud detection: **99,000 non-fraud cases** vs. **1,000 fraud cases**.
- Medical diagnosis: **95% healthy patients** vs. **5% with disease**.
- Customer churn prediction: **90% retained** vs. **10% churned**.

⍰ **Challenges:**

- The model is biased towards the majority class.
- Poor performance in minority class predictions.
- Accuracy can be misleading (e.g., a model predicting "not fraud" 99% of the time but failing to detect fraud).

Main Example:-

Transaction ID	Amount ($)	Location	Fraud (Target Label)
1	50	New York	**No (0)**
2	200	Chicago	**No (0)**
3	5000	Los Angeles	**Yes (1)**
4	20	Miami	**No (0)**
5	150	Boston	**No (0)**
6	10000	Houston	**Yes (1)**
7	30	Denver	**No (0)**
8	250	Seattle	**No (0)**

Solutions to Handle Unbalanced Data:

- **Resampling Techniques**:

    - **Oversampling** the minority class (e.g., SMOTE – Synthetic Minority Over-sampling Technique).
    - **Undersampling** the majority class to reduce imbalance.
- **Class Weight Adjustment**:
    - Assigning higher weights to the minority class during model training.
- **Anomaly Detection Approaches**:
    - Treating the minority class as an anomaly (useful in fraud detection).
- **Performance Metrics Instead of Accuracy**:
    - **Precision, Recall, F1-score** instead of accuracy.
    - **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)** to measure model performance.

ASSIGNMENT-2

Business Understanding:-

It involves clearly defining the business problem, objectives, and success criteria before diving into data collection and modeling.

Business problem

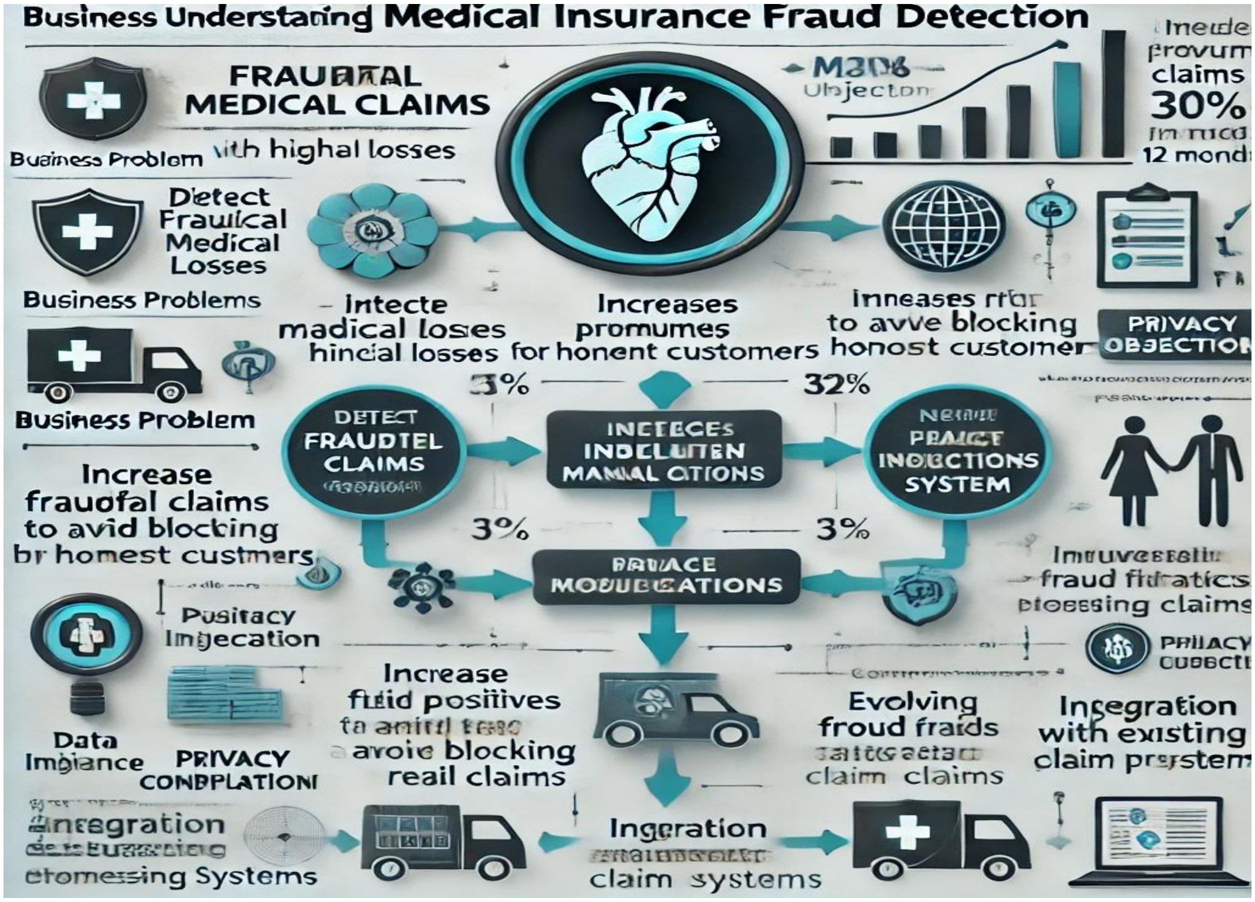
Business Objective

Business Constraint

These 3 are play most crucial role for Business understanding

Example:- medical insurance fraud detection

Business Problem: Medical insurance fraud is a significant issue, leading to financial losses and impacting the trust in the system.
Business Objective: Automatically detect fraud with high accuracy, reduce fraudulent payouts by **30%**, and improve claims processing efficiency.
Business Constraints: Data imbalance, privacy regulations, evolving fraud tactics, need for model transparency, and integration with existing workflows.


Business Understaring Medical Insurance Fraud Detection
FRAUBITAL MEDICAL CLAIMS
Business Problem
vith highal losses
M3D6
Ubjecton
Ineude provum claims 30% in micd 12 mond
Dietect Fraulical Medical Losses
Business Problems
- Intecte madical loses hincial losses for honent customers
Increases promumes
Inneases rfor to avve blocking honost customer
PRIVACY OBJECTION
3%
32%
Business Problem
Increase fraufal claims to avid blocking br homest custmers
DETECT FRAUDTEL CLAIMS
INECECES INDELUTEN MANAL OTIONS
NEBIE PRAICT INOBCTIONS SYSTEM
3%
3%
PRIVACE MOUDBCATIONS
Imuvessiti fraud fidratics ptoessing claims
Pusitacy Ingecation
PRIIACY OUBSCTI
Data Imgiance
PRIVACY CONDPLATIONI
Increase fuid positives avoie blocking reail claims
Evolving froud fraids claim claims
Ingegration with existing claim prystem
integration etomessing Systems
Ingration claim systems

ASSIGNMENT-3

Confusion Matrix:-

A **confusion matrix** is a performance measurement tool used in classification problems to evaluate the accuracy of a model.

It compares the predicted classifications against the actual ground truth values.

The matrix provides insights into how well the model is performing by showing not only correct predictions but also where it is making errors.

Structure of a Confusion Matrix:-

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

Here,

- **True Positive (TP):** Correctly predicted positive cases (the model predicted positive, and it was actually positive).
- **True Negative (TN):** Correctly predicted negative cases (the model predicted negative, and it was actually negative).
- **False Positive (FP):** Incorrectly predicted positive cases (the model predicted positive, but it was actually negative).
- **False Negative (FN):** Incorrectly predicted negative cases (the model predicted negative, but it was actually positive).

Example: Medical Insurance Fraud Detection

Imagine we have a model that detects **fraudulent** vs. **non-fraudulent** insurance claims. After testing the model on a dataset, we obtain the following results:

- **Predicted Fraud (Yes)**
- **Predicted Not Fraud (No)**

Confusion Matrix for the Model:

Actual	**Predicted Fraud (Yes)**	**Predicted Not Fraud (No)**
Actual Fraud (Yes)	85 (TP)	15 (FN)
Actual Not Fraud (No)	20 (FP)	80 (TN)