EDA and Project Explanation

1. False Positive (FP):

- Definition: A false positive occurs when the model incorrectly predicts the positive class (1) for an instance that actually belongs to the negative class (0).
- Example: In a medical diagnosis scenario, a false positive would be when the model predica patient has a disease (positive class) when they do not actually have it (negative class).

2. False Negative (FN):

- Definition: A false negative occurs when the model incorrectly predicts the negative class (0) for an instance that actually belongs to the positive class (1).
- Example: Continuing with the medical diagnosis example, a false negative would be when the model predicts a patient does not have a disease (negative class) when they actually do have it (positive class).

1. Handling many Null values in almost all columns:

This suggests that the dataset contains a substantial amount of missing data, which needs to be addressed before building a predictive model. Various strategies such as imputation, deletion, or advanced techniques like multiple imputation may need to be employed to handle these missing values effectively.

2. No low-latency requirement:

Low-latency requirements typically mean that predictions need to be made quickly, often in real-time or near-real-time. In this case, since there's no such requirement, the focus can be on building accurate models even if they take longer to train or make predictions.

3. Interpretability is not important:

This implies that the primary objective is predictive accuracy rather than understanding the underlying factors driving the predictions. Models

that are highly accurate but complex, such as deep neural networks or ensemble methods, may be preferred over simpler, more interpretable models like logistic regression or decision trees.

4. Misclassification leads to unnecessary repair costs:

This indicates that the cost of misclassifying instances, particularly false positives and false negatives, is associated with repair costs. For example, in a manufacturing setting, misclassifying a defective product as non-defective (false negative) could lead to costly repairs or replacements down the line. Similarly, incorrectly classifying a non-defective product as defective (false positive) could result in unnecessary repair costs.

KNN imputer

The KNNImputer is a method used for imputing missing values in a dataset based on the k- nearest neighbors approach. It replaces missing values in a feature with the mean value of the feature's k-nearest neighbors. Here's how it works:

1. For each sample with missing values, the algorithm finds its k-nearest neighbors based on other samples with non-missing values.
2. It calculates the mean value of the feature among these k-nearest neighbors.
3. This mean value is then used to replace the missing value in the original sample.

I. StandardScaler:

- Standardizes features by removing the mean and scaling to unit variance.
- Sensitive to outliers because it computes the mean and standard deviation, which can be influenced by extreme values.
- Works well when the data is normally distributed and does not contain many outliers.

II. RobustScaler:

- Scales features using statistics that are robust to outliers.
- It uses the median and interquartile range (IQR) instead of the mean and standard deviation.

- More suitable for data with outliers or skewed distributions because it's less affected by extreme values.
- It's not sensitive to the presence of outliers as it focuses on the median and quartiles

III. MinMaxScaler:

- Scales features to a specified range, typically between 0 and 1.
- It's sensitive to outliers because the scaling is based on the minimum and maximum values of each feature.
- Preserves the shape of the original distribution, but it may not handle outliers well.

IV. RobustScaler:

- Scales features using statistics that are robust to outliers, such as the median and interquartile range (IQR).
- It's less affected by outliers compared to MinMaxScaler because it uses more robust statistics.
- It's suitable for data with outliers or skewed distributions because it adjusts the scaling based on the median and quartiles.

V. SMOTE (Synthetic Minority Over-sampling Technique):

- SMOTE works by creating synthetic samples from the minority class (the class with fewer instances) rather than duplicating them. It randomly selects a minority class instance and finds its k nearest minority class neighbors. Then, it randomly selects one of these neighbors and creates a synthetic instance along the line segment joining the two points in the feature space. This process helps to balance the class distribution by increasing the number of minority class instances.

VI. Tomek links:

- Tomek links are pairs of instances from different classes that are very close to each other in the feature space. By removing such instances, Tomek links can help clarify the decision boundary between classes. In the context of SMOTE, applying Tomek links after generating synthetic samples helps to improve the quality of the synthetic data by removing noisy or ambiguous instances.

VII. SMOTE+TOMEK:

- SMOTE+TOMEK is a combination of SMOTE oversampling and Tomek links undersampling. It first applies SMOTE to oversample the minority class, and then applies Tomek links to remove potentially noisy or ambiguous instances from both the minority and majority classes. This combined approach aims to create a more balanced and clearer representation of the dataset

Steps you need to follow :

- Reading the data in python
- Defining the problem statement
- Identifying the Target variable
- Looking at the distribution of Target variable
- Basic Data exploration
- Rejecting useless columns
- Visual Exploratory Data Analysis for data distribution (Histogram and Bar Charts)
- Feature Selection based on data distribution
- Outlier treatment
- Missing Values treatment
- Visual correlation analysis
- Statistical correlation analysis (Feature Selection)
- Converting data to numeric for ML
- Sampling and K-fold cross validation
- Trying multiple Regression algorithms
- Selecting the best Model