# Mention the Data Preprocessing steps you have implemented and why you required that in this project?

The preprocessing pipeline used for data imputation, encoding categorical variables, and scaling numerical variables. Below, we outline the key components and their importance in the data preparation process.

1. **Specifying Columns for Encoding and Scaling**
   * **Categorical Columns:** These columns will be one-hot encoded to represent categorical features in a machine-readable format. Examples include job, marital, education, etc.
   * **Numerical Columns:** These columns, such as age, balance, duration, etc., will be scaled to normalize their values.
   * **Why This Step is Required:** The distinction between categorical and numerical columns is crucial for preprocessing. Different types of features need appropriate transformations—categorical columns are one-hot encoded to convert non-numeric data into binary vectors, while numerical columns are scaled to bring their values to a common range, which helps improve the model's performance
2. **Defining the Preprocessor**
   * **Purpose:** The preprocessor combines all preprocessing steps into a single pipeline using ColumnTransformer from sklearn.compose.
   * **ColumnTransformer:** Allows us to apply different transformations to different columns in the dataset.
     + **One-Hot Encoding:** Encodes categorical columns into dummy variables to convert non-numeric data into binary vectors. It ensures unknown categories are ignored, avoiding errors, and uses 32-bit integers for the output.
     + **Min-Max Scaling:** Scales numerical columns to bring values to a range between 0 and 1, helping normalize the data.
   * **Passthrough Remainder:** Keeps any columns not specified in the transformations unchanged.
   * **Why This Step is Required:** One-hot encoding and scaling are vital for preparing the data for machine learning. Encoding helps convert categorical data into a format that models can understand, while scaling ensures that numerical features contribute equally, preventing bias due to different magnitudes.

## Summary

The preprocessing pipeline defined here includes one-hot encoding for categorical features and min-max scaling for numerical features. These transformations help create a standardized format for machine learning models, ensuring effective data representation during training and testing.

The next step involves applying this preprocessor to the dataset using a model pipeline to ensure consistency during both training and inference phases.

# Explain the Initially trained model and retraining process you have used

I experimented with SelectKBest and PCA for feature selection and dimensionality reduction, respectively, on different classification models: Logistic Regression, RandomForestClassifier, and BaggingClassifier. Here are my key observations:

1. **Initial Training**:
   * During the initial training phase, I used all available features for each of the models to establish a performance baseline.
   * The models included **Logistic Regression**, **RandomForestClassifier**, and **BaggingClassifier**. Using all features allowed me to understand the maximum predictive capability of the models without any feature selection or dimensionality reduction.
2. **Retraining with Feature Selection**:
   * After the initial training, I applied **SelectKBest** with different values of k (ranging from **1 to 34**) to identify the most important features and reduce dimensionality.
   * With **k=3**, the performance of the models showed almost no loss compared to using the full feature set. This indicated that only a few features were highly significant, allowing for simpler models without compromising accuracy.
   * I also applied **PCA** (Principal Component Analysis) during retraining to reduce dimensionality by transforming features into principal components. However, PCA did not lead to any significant performance improvement and instead increased computational time.
3. **Overall Conclusion:**
   * **SelectKBest** with a lower k value provided the most significant improvement during the retraining phase, but it did not yield significant results on the test set.

# Mention the references you have used for these tasks

ChatGPT