Assignment: Classification with Machine Learning Models

Problem Statement

The goal of this assignment is to apply **multiple classification algorithms** to real-world datasets. You will build, evaluate, and compare models such as **Decision Tree, Random Forest, AdaBoost, XGBoost, and CatBoost**.

You are required to perform **data exploration, preprocessing, model building, and optimization** to understand how classification models behave on different datasets.

Dataset Links

1. Telco Customer Churn – Kaggle
2. HR Analytics (Employee Attrition) – Kaggle
3. Stroke Prediction – Kaggle

Assignment Guidelines

1. Data Understanding

- Load the dataset and display the first few rows.
- Identify the **input features** and the **target variable**.
- Check data types (numerical vs categorical).
- Check for **missing values and duplicates**.

2. Exploratory Data Analysis (EDA)

- Plot the distribution of the target variable.
- Visualize relationships between features and the target (e.g., Age vs Survival, MonthlyCharges vs Churn).
- Compare categories (e.g., Gender, Department, Smoking Status).
- Create a **correlation heatmap** for numerical features.

3. Data Preprocessing

- Encode categorical variables.
- Scale numerical features if required.
- Handle missing values appropriately.
- Split into **training and testing sets**.

4. Model Building (Apply All Classifiers)

You must apply the following classifiers **one by one**:

1. Decision Tree
2. Random Forest
3. AdaBoost
4. XGBoost
5. CatBoost

For each classifier:

- Train the model on training data.
- Evaluate on testing data using:
    - **Accuracy, Precision, Recall, F1-score**

  - **Confusion Matrix**

5. Model Optimization

- Perform **hyperparameter tuning** for at least 2 models (e.g., Random Forest & XGBoost).
- Compare tuned performance vs default.
- Discuss **overfitting/underfitting observations**.

6. Model Evaluation and Comparison

- Compare the performance of all 5 models in a **summary table**.
- Identify the **best-performing model**.
- Discuss which features are most important (Feature Importance plots for tree-based models).