ML Problem Statements

1. IRIS Dataset

Problem Statement

The objective is to develop a machine learning model that can accurately classify the species of an iris flower based on its physical characteristics. The dataset provides measurements of sepal length, sepal width, petal length, and petal width for each flower, along with its species. The three species in the dataset are Setosa, Versicolor, and Virginica.

Goal  
To build a classification model that can predict the species of an iris flower given its sepal and petal measurements.

Dataset  
The dataset used is the Iris dataset, which consists of 150 samples with 4 features (sepal length, sepal width, petal length, petal width) and a target label indicating the species of the flower.

Data Description

* Sepal Length: The length of the sepal in centimeters.
* Sepal Width: The width of the sepal in centimeters.
* Petal Length: The length of the petal in centimeters.
* Petal Width: The width of the petal in centimeters.
* Species: The class label, which represents the species of the iris flower (Setosa, Versicolor, Virginica).

Machine Learning Techniques  
We will explore multiple classification algorithms to develop the model, including:

1. Logistic Regression
2. Decision Trees
3. Random Forests
4. Support Vector Machine (SVM)

Evaluation Metrics  
The model's performance will be evaluated using:

* Accuracy: The percentage of correctly classified samples.
* Precision: The ratio of correctly predicted positive observations to the total predicted positives.
* Recall: The ratio of correctly predicted positive observations to all observations in the actual class.
* F1-Score: The weighted average of Precision and Recall, which accounts for both false positives and false negatives.

Project Requirements

* Data Preprocessing: Handling any missing data (if applicable), feature scaling, and encoding the target variable for model compatibility.
* Model Training: Training multiple classification algorithms and selecting the best-performing model.
* Model Evaluation: Evaluating the model's performance on a test set and tuning hyperparameters for optimal results.
* Interpretability: Providing insights into which features contribute the most to the model's predictions.

Deliverable The final deliverable is a classification model capable of predicting the species of an iris flower with high accuracy, along with a report summarizing model performance and insights.

2. Tips Dataset

Problem Statement

The goal is to develop a predictive model that estimates the tip amount a customer is likely to leave based on various features related to the dining experience. Using the Tips dataset, we aim to understand the factors that influence tipping behavior and predict the tip amount based on features such as the total bill, the customer’s gender, smoking status, day and time of the visit, and the size of the dining party.

Objective  
To build a regression model that can predict the tip amount given relevant characteristics of the dining experience.

Dataset  
The dataset is the Tips dataset, which contains information on bills and tips from a restaurant. Each record includes:

* Total Bill: The total amount of the bill in dollars.
* Tip: The tip amount in dollars (target variable).
* Gender: Gender of the customer paying the bill.
* Smoking Status: Whether the table was in a smoking or non-smoking section.
* Day: Day of the week when the meal took place.
* Time: Time of day (Lunch or Dinner).
* Party Size: The number of people in the dining party.

Machine Learning Techniques  
To predict the tip amount, we will apply regression analysis using:

1. Linear Regression: To examine a linear relationship between the features and the tip amount.
2. Polynomial Regression: To capture any potential non-linear relationships in the data.

Evaluation Metrics  
We will evaluate the performance of the model using:

* Mean Squared Error (MSE): Measures the average squared difference between predicted and actual tip amounts, penalizing larger errors.
* R-Squared (R²): Indicates the proportion of variance in the tip amount that is predictable from the input features.

Project Requirements

* Data Preprocessing: Handling any missing values, encoding categorical variables (e.g., gender, smoking status), and normalizing or scaling numerical features if needed.
* Feature Selection: Identifying the most relevant features for predicting tip amount to improve model performance and interpretability.
* Model Training: Building and training regression models to find the best fit for the data.
* Hyperparameter Tuning: Adjusting model parameters to optimize predictive accuracy.

Deliverable The final deliverable will be a regression model that can accurately predict the tip amount based on the given features, along with a report detailing model performance, feature importance, and insights into tipping behavior.

3. Titanic Datasets

### Titanic Dataset - Detailed Problem Statement

#### Background

The Titanic dataset consists of data about the passengers aboard the ill-fated RMS Titanic, which sank on its maiden voyage in 1912. The dataset provides details about passengers, including demographic information (e.g., age, gender), ticket-related details (e.g., ticket class, fare), and whether they survived the disaster. This dataset has been widely used in machine learning for classification tasks, especially to build predictive models for survival prediction based on various passenger features.

#### Problem Definition

The goal is to build a machine learning classification model that can predict whether a passenger survived the Titanic disaster based on a range of features. By predicting survival, the model can provide insights into factors influencing survival chances, which can be critical in understanding patterns of human behavior during disasters.

#### Input Features

The dataset contains multiple features about each passenger:

* PassengerId: A unique identifier for each passenger.
* Pclass: The passenger's ticket class (1st, 2nd, or 3rd class).
* Name: Name of the passenger.
* Sex: Gender of the passenger (male/female).
* Age: Age of the passenger (in years).
* SibSp: Number of siblings or spouses aboard the Titanic.
* Parch: Number of parents or children aboard the Titanic.
* Ticket: The ticket number.
* Fare: The fare the passenger paid for the ticket.
* Cabin: The cabin where the passenger stayed (this data is mostly missing).
* Embarked: The port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton).
* Survived: The target variable, which indicates whether the passenger survived (1) or did not survive (0).

#### Objective

The task is to predict the target variable Survived, which is binary (0 or 1), based on the other features. The prediction model needs to classify a passenger into one of two categories:

1. Survived (1): The passenger survived the disaster.
2. Did Not Survive (0): The passenger did not survive the disaster.

The classification model will help in understanding what factors may have contributed to the chances of survival on the Titanic. The primary focus will be on leveraging attributes like gender, age, passenger class, and fare, as these are often considered important in survival analysis.

#### Approach

1. Data Preprocessing:
   * Handling Missing Data: Some features, such as age and cabin, may have missing values, so we will either impute missing values (for age) or drop columns with excessive missing data (for cabin).
   * Feature Encoding: The categorical variables (e.g., Sex, Embarked, Pclass) need to be encoded to numerical values, using techniques like one-hot encoding or label encoding.
   * Feature Scaling: If necessary, the features (e.g., Age, Fare) will be scaled or normalized to ensure that models like Logistic Regression or SVM can perform well.
2. Modeling Techniques: We will try several classification algorithms and compare their performance:
   * Logistic Regression: A simple and interpretable linear model suitable for binary classification.
   * Decision Trees: A non-linear model that splits data based on feature values and can capture complex relationships.
   * Random Forests: An ensemble learning method that uses multiple decision trees to improve accuracy and reduce overfitting.
   * Support Vector Machine (SVM): A powerful classifier that finds an optimal hyperplane to separate classes in a higher-dimensional space.
3. Evaluation Metrics: The model will be evaluated using multiple performance metrics to ensure a comprehensive understanding of its capabilities:
   * Accuracy: The percentage of correctly predicted instances.
   * Confusion Matrix: A table that helps in understanding the types of errors made by the model, showing True Positives, True Negatives, False Positives, and False Negatives.
   * ROC-AUC Score: The Area Under the Receiver Operating Characteristic curve, which provides an aggregate measure of the model's ability to discriminate between classes across different thresholds.
   * Precision, Recall, and F1-Score: These metrics will help evaluate the model's performance, especially in cases where the class distribution is imbalanced (e.g., there are more non-survivors than survivors).
4. Model Tuning:
   * Hyperparameter Tuning: Techniques such as grid search or random search will be employed to find the optimal hyperparameters for each model to maximize performance.
   * Cross-validation: To reduce the risk of overfitting and ensure that the model generalizes well to unseen data, k-fold cross-validation will be used.

#### Challenges

* Missing Data: Some of the features, especially Age and Cabin, have missing values that need to be handled carefully.
* Imbalanced Classes: The Titanic dataset may have an imbalance between the number of survivors and non-survivors, which can affect model performance.
* Feature Engineering: Identifying and creating new features (such as family size by combining SibSp and Parch) could help improve model performance.

#### Deliverables

The final deliverable will be:

1. A well-trained classification model capable of predicting whether a passenger survived the Titanic disaster.
2. A report summarizing the model’s performance, the significance of different features, and insights gained from the analysis (e.g., how age, gender, or class influenced survival chances).
3. A final model with optimized hyperparameters ready for deployment or further analysis.