### **Exercise 1 : Defining the Problem and Data Collection for Loan Default Prediction**

### **Problem Statement**

The goal of this project is to develop a predictive model to identify potential loan defaults by analyzing applicant data and loan characteristics. Accurate prediction of defaults will enable financial institutions to minimize risk and optimize lending practices.

### **Data Types Needed**

* **Personal Details:** Age, income, employment status, marital status, etc.
* **Credit Scores:** Historical credit ratings, credit utilization, and past defaults.
* **Loan Details:** Loan amount, interest rates, loan type, and term duration.
* **Repayment History:** Payment dates, amounts paid, missed payments, and outstanding balances.

### **Data Sources**

* **Financial Institution Records:** Internal data on loans and repayment histories.
* **Credit Bureaus:** Credit scores and historical credit behavior.
* **Government Databases:** Socioeconomic data for contextual analysis.
* **Open Data Portals:** General economic indicators that may influence default risk.

### **Exercise 2 : Feature Selection and Model Choice for Loan Default Prediction**

### **Relevant Features for Predicting Loan Defaults**

1. **Credit Score:** A direct indicator of an applicant's creditworthiness and historical financial behavior.
2. **Repayment History:** Tracks missed or late payments, which are strong predictors of future defaults.
3. **Income:** Determines the applicant's ability to repay the loan based on their financial stability.
4. **Loan Amount:** Larger loan amounts may increase default risk, especially if disproportionate to income.
5. **Debt-to-Income Ratio:** Combines income and existing debts, showing the applicant's financial burden.
6. **Employment Status:** Provides insight into job stability and future income potential.

### **Exercise 3 : Training, Evaluating, and Optimizing the Model.**

**Split Data:** Train/test split.

**Train Model:** Use training data.

**Evaluate Metrics:** Use accuracy, precision, recall, F1-score on test data.

**Optimize:** Tune hyperparameters.

### **Exercise 4 : Designing Machine Learning Solutions for Specific Problems**

**Predicting Stock Prices:**

* **Type:** **Supervised Learning (Regression)**
* **Explanation:** Stock prices are continuous, and the model needs to predict future values based on historical data.

**Organizing a Library of Books:**

* **Type:** **Unsupervised Learning (Clustering)**
* **Explanation:** Books can be grouped into genres or categories based on features like content, author, or keywords without predefined labels.

**Programming a Robot to Navigate a Maze:**

* **Type:** **Reinforcement Learning**
* **Explanation:** The robot learns optimal actions through trial and error to find the shortest path, maximizing rewards for reaching the goal efficiently.

### **Exercise 5 : Designing an Evaluation Strategy for Different ML Models**

### **1. Supervised Learning Model: Decision Tree Classifier**

**Evaluation Strategy:**

* **Metrics:** Accuracy, precision, recall, F1-score, AUC-ROC.
  + **Accuracy:** Overall correctness but may be skewed by class imbalance.
  + **Precision & Recall:** Important for understanding the trade-off between false positives and false negatives, particularly in high-stakes applications.
  + **F1-Score:** Balances precision and recall, useful when there's a trade-off.
  + **AUC-ROC:** Evaluates the model's ability to distinguish between classes across various thresholds.
* **Methods:** Cross-validation and ROC curves.
  + **Cross-Validation:** Ensures robustness by testing the model on multiple subsets of data, mitigating the risk of overfitting.
  + **ROC Curves:** Visualizes performance across different thresholds, helping select the optimal decision boundary.

**Challenges & Limitations:**

* **Imbalanced Datasets:** Accuracy can be misleading; precision and recall become more critical.
* **Overfitting:** Cross-validation helps, but complex models may still overfit, requiring careful tuning.

### **2. Unsupervised Learning Model: K-Means Clustering**

**Assessment Strategy:**

* **Techniques:** Silhouette score, elbow method, and cluster validation metrics.
  + **Silhouette Score:** Measures how well samples are clustered, with higher values indicating better-defined clusters.
  + **Elbow Method:** Helps determine the optimal number of clusters by identifying the point where adding more clusters does not significantly improve the model.
  + **Cluster Validation Metrics:** Such as Davies-Bouldin Index, to evaluate cluster compactness and separation.

**Challenges & Limitations:**

* **Subjectivity:** Determining the "correct" number of clusters can be subjective and may require domain knowledge.
* **Cluster Interpretability:** Even with good validation scores, clusters may not be meaningful or actionable without clear interpretability.

### **3. Reinforcement Learning Model: Q-Learning**

**Evaluation Strategy:**

* **Measures:** Cumulative reward, convergence, exploration vs. exploitation balance.
  + **Cumulative Reward:** Indicates the total success of the agent over time, providing insight into overall performance.
  + **Convergence:** Determines if the model consistently reaches an optimal policy, signifying learning stability.
  + **Exploration vs. Exploitation Balance:** Evaluates how well the agent balances trying new actions (exploration) with leveraging known rewards (exploitation).

**Challenges & Limitations:**

* **Convergence Issues:** RL models may take a long time to converge, or they may converge to suboptimal solutions.
* **Exploration vs. Exploitation:** Striking the right balance is challenging; too much exploration can lead to inefficiency, while too much exploitation may prevent discovering better strategies.