**Step 1: Check the Data’s Health**

Before doing anything fancy, we check if our data is clean:

* Are there any missing values?
* Does every column have the right type of information?

Think of it like checking ingredients before cooking—you don’t want rotten tomatoes in your salad.

**Step 2: Understand What Might Make People Quit (EDA)**

We explore the data visually to look for clues. This helps answer: *Why are employees leaving?*

**🔹 Look at correlations:**

* Use a heatmap to see which factors are strongly connected.

Example: Low satisfaction might be highly related to leaving.

**🔹 Look at key feature distributions:**

* How satisfied are employees?
* How were they rated in their last evaluation?
* How many hours do they work per month?

**🔹 Compare people who left vs. those who stayed:**

* Use a bar graph to compare how many projects each group worked on.

This helps HR understand what kind of work environment causes burnout or boredom.

**Step 3: Group Similar Quitters (Clustering)**

We look *only* at the employees who left and try to find patterns among them:

* Group them based on satisfaction and evaluation scores using something called **K-means clustering**.
* We aim to find 3 distinct types (or “clusters”) of leavers.

This tells us, for example, if some people left because they were too stressed, others left due to boredom, etc.

**Step 4: Balance the Data**

Normally, more people *stay* than *leave*, which creates an imbalance that confuses machine learning models.

So, we use a trick called **SMOTE** to "imagine" new data points for people who left, making the classes (stay/left) balanced.

But before that:

* We convert department names and salary levels (text data) into numbers so the model can understand them.
* We split the data into a training set (to teach the model) and a test set (to check if the model learned well).

**Step 5: Train Three Prediction Models**

Now we teach 3 different types of ML models to predict if someone will leave:

1. Logistic Regression – simple and interpretable.
2. Random Forest – uses many decision trees.
3. Gradient Boosting – a powerful ensemble technique.

We test each model using a method called **5-fold cross-validation**, which helps make sure the model performs well on different data subsets.

**Step 6: Pick the Best Model and Justify**

We compare the 3 models using:

* **ROC Curve & AUC Score**: Tells us how well the model separates "stay" vs. "leave".
* **Confusion Matrix**: Shows how many employees were correctly/incorrectly predicted to stay/leave.
* **Precision vs Recall**:
  + Use **recall** if you care about catching as many leavers as possible.
  + Use **precision** if you want to avoid false alarms.

**Step 7: Create a Risk Dashboard and Retention Plan**

Now that we have a good model, we:

* Predict the probability that each employee might leave.
* Divide them into 4 zones based on the score:
  + **Green (Safe)**: < 20% chance of leaving
  + **Yellow (Low Risk)**: 20–60%
  + **Orange (Medium Risk)**: 60–90%
  + **Red (High Risk)**: > 90%

This is like a weather forecast for HR—helping them take action **before** the storm (resignation) hits.

Based on which zone employees fall into, HR can design personalized strategies like:

* Red zone: Give incentives, check workload, offer flexibility.
* Yellow zone: Engage with them more, give recognition.
* Green zone: Keep doing what’s working.

**🧠 End Result**

You’ve created a data-driven system that:

* Predicts who’s likely to leave,
* Understands why they leave,
* Groups them into actionable buckets, and
* Helps HR save valuable talent.