**Now I will talk about how we train our Machine Learning Model**

**There are four steps, data preprocessing, Model Training, Model Evaluation and Feature Importances**

1. **Data Preprocessing:**

**Encoding Categorical Variables:**

We encode Categorical variables using label encoding to convert them into numerical format. This step is necessary as Random Forest models require numerical inputs.

**Balancing the dataset:**

Our dataset has an imbalance between the Attrition Yes and Attrition No classes with 20% Yes and 80% No. This imbalance can lead to a poor performance in identifying the minority class.

Therefore, we use SMOTE to generate synthetic examples of the minority class to balance the dataset, this allows the model to learn better from the minority class data, improving its ability to correctly predict the minority class.

**Train-Test Split:**

We split the dataset into training and testing sets with 70 –30 portion. This split ensures that the model is trained on a large portion of the data while leaving enough data to test its performance.

**2. Model Training**

**Model Choice:**

We use Random Forest to train our model

Random Forest is an ensemble learning method used for classification and regression tasks. It works by constructing multiple decision trees during training and outputting the mode (for classification) or mean (for regression) of the individual trees' predictions. Each tree in a Random Forest is trained on a different subset of the data, using a random selection of features.

We select Random Forest to train our model because it can work well with both numerical and categorical data.

It also provides insights into feature importance, helping us understand which factors contribute the most to employee attrition, which is valuable for HR context.

**Hyperparameter Tuning:**

To optimize the model, we use GridSearchCV to tune hyperparameters. The hyperparameters grid for Random Forest includes the number of trees, maximum tree depth, minimum samples required to split a node, and minimum samples required at a leaf node.

GridSearchCV will perform a comprehensive search by training the model on every possible combination of hyperparameters specified in the grid

GridSearchCV then performs cross-validation on each combination of hyperparameters to evaluate model performance.

The model with the best cross-validated performance is selected as the final model.

**4. Evaluate the model:**

The hyperparameter tuning process further improves the model's performance, from 89% to 91% accuracy score. This suggests that the tuning process was successful in optimizing the model for better prediction accuracy.

Beside the impressive accuracy 91% of the model, we also notice that both classes have an F1 score of 91%, indicating that the model is equally effective at both identifying employees who will stay and employees who will leave, making it valuable for HR decision-making and strategic planning.

**5. Feature Importances:**

The feature importances showing that financial incentives (like salary and stock options) and working environment (like job satisfaction and environment satisfaction) are the strongest factors in retaining employees. This suggests that to reduce attrition, companies should focus on offering competitive compensation packages and ensuring a positive working environment.

Now I will handover to Rekka for the model deployment