Problem Statement: Predicting Purchase Behaviour Based on Demographic and Financial Data

Objective:

To build a machine learning classification model that can predict whether a customer will purchase a product based on their age, estimated salary, and gender. This model will help in understanding customer purchase behaviour and targeting potential buyers more effectively.

Business Context:

A retail company wants to increase its sales by identifying potential customers who are more likely to purchase its products. By analysing the existing customer data, the company aims to predict future purchase behaviour and tailor its marketing strategies accordingly.

Data Description:

The dataset contains the following features:

- **Age**: The age of the customer.
- **EstimatedSalary**: The estimated annual salary of the customer.
- **Gender_Male**: A binary variable indicating the gender of the customer (1 if male, 0 if female).

The target variable is:

- **Purchased**: A binary variable indicating whether the customer purchased the product (1 if purchased, 0 if not purchased).

Problem:

Given the demographic and financial information of a customer, predict whether they will purchase a product.

Confusion Matrix:

The confusion matrix represents the performance of a classification model. Here's the confusion matrix for the given data:

```
[[71,  8],

 [ 3, 38]]
```

Random Forest (RF) Classification:

To address this problem, we will utilize the Random Forest (RF) classification algorithm. Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

1. **Precision**:
   - Precision = TP / (TP + FP)
2. **Recall (Sensitivity)**:
   - Recall = TP / (TP + FN)
3. **F1-Score**:
   - F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
4. **Accuracy**:
   - Accuracy = (TP + TN) / (TP + TN + FP + FN)
5. **Macro Average**:
   - Macro Average = (Metric for Class 0 + Metric for Class 1) / 2
6. **Weighted Average**:
   - Weighted Average = (Metric for Class 0 × Support for Class 0 + Metric for Class 1 × Support for Class 1) / Total Support

Given Data:

- True Positives (TP) for Class 0 = 71
- False Positives (FP) for Class 0 = 3
- False Negatives (FN) for Class 0 = 8
- True Negatives (TN) for Class 0 = 38
- True Positives (TP) for Class 1 = 38
- False Positives (FP) for Class 1 = 8
- False Negatives (FN) for Class 1 = 3
- True Negatives (TN) for Class 1 = 71

Evaluation Metrics:

- **Precision:**
  - Precision = TP / (TP + FP)
  - **Precision for "not purchased" (class 0):** Precision 0 = 71 / (71 + 3) = 71 / 74 ≈ 0.96
  - **Precision for "purchased" (class 1):** Precision 1 = 38 / (38 + 8) = 38 / 46 ≈ 0.83
- **Recall (Sensitivity):**
  - Recall = TP / (TP + FN)
  - **Recall for "not purchased" (class 0):** Recall 0 = 71 / (71 + 8) = 71 / 79 ≈ 0.90
  - **Recall for "purchased" (class 1):** Recall 1 = 38 / (38 + 3) = 38 / 41 ≈ 0.93
- **F1-Score:**
  - F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
  - **F1-Score for "not purchased" (class 0):** F1-Score 0 = 2 × (0.96 × 0.90) / (0.96 + 0.90) ≈ 0.93
  - **F1-Score for "purchased" (class 1):** F1-Score 1 = 2 × (0.83 × 0.93) / (0.83 + 0.93) ≈ 0.87
- **Accuracy:**
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)
  - **Accuracy:** Accuracy = (71 + 38) / (71 + 38 + 3 + 8) = 109 / 120 ≈ 0.91

- **Macro Average:**
    - Macro Average = (Metric for Class 0 + Metric for Class 1) / 2
    - **Macro Precision:** (0.96 + 0.83) / 2 ≈ 0.89
    - **Macro Recall:** (0.90 + 0.93) / 2 ≈ 0.91
    - **Macro F1-Score:** (0.93 + 0.87) / 2 ≈ 0.90
- **Weighted Average:**
    - Weighted Average = (Metric for Class 0 × Support for Class 0 + Metric for Class 1 × Support for Class 1) / Total Support
    - **Weighted Precision:** (0.96 × 79 + 0.83 × 41) / 120 ≈ 0.91
    - **Weighted Recall:** (0.90 × 79 + 0.93 × 41) / 120 ≈ 0.91
    - **Weighted F1-Score:** (0.93 × 79 + 0.87 × 41) / 120 ≈ 0.91

Classification Report Interpretation - Random Forest (RF)

The classification report provides valuable insights into the performance of the Random Forest (RF) classification model. Here's an interpretation of the metrics:

Precision:

Precision measures the proportion of true positive predictions among all positive predictions made by the model.

- **Precision for Class 0 ("not purchased"):** Precision is 0.96, indicating that out of all instances predicted as "not purchased" by the model, 96% were actually "not purchased". This suggests a high level of accuracy in identifying customers who did not make a purchase.
- **Precision for Class 1 ("purchased"):** Precision is 0.83, meaning that out of all instances predicted as "purchased" by the model, 83% were actually "purchased". While slightly lower than for Class 0, this still represents a relatively high precision in identifying customers who made a purchase.

Recall (Sensitivity):

Recall measures the proportion of actual positives that were correctly predicted by the model.

- **Recall for Class 0 ("not purchased"):** Recall is 0.90, indicating that the model correctly identified 90% of the actual "not purchased" instances. This suggests that the model effectively captures most of the instances where customers did not make a purchase.
- **Recall for Class 1 ("purchased"):** Recall is 0.93, indicating that the model captured 93% of the actual "purchased" instances. This indicates a high sensitivity in identifying customers who made a purchase.

F1-Score:

The F1-score is the harmonic mean of precision and recall and provides a balance between the two metrics.

- **F1-Score for Class 0 ("not purchased"):** The F1-score is 0.93, reflecting the balance between precision and recall for "not purchased". This indicates a good overall performance in predicting instances where customers did not make a purchase.
- **F1-Score for Class 1 ("purchased"):** The F1-score is 0.87, indicating a good balance between precision and recall for "purchased". This suggests a solid overall performance in predicting instances where customers made a purchase.

Support:

Support refers to the number of instances in each class.

- **Support for Class 0 ("not purchased"):** There are 79 instances of "not purchased" in the test set.
- **Support for Class 1 ("purchased"):** There are 41 instances of "purchased" in the test set.

Accuracy:

Accuracy measures the overall correctness of the model's predictions across all classes.

- **Accuracy:** The overall accuracy is 0.91, meaning that the model predicted 91% of instances correctly across both classes. This indicates a high level of overall predictive accuracy.

Conclusion:

The Random Forest (RF) classification model demonstrates strong performance in predicting customer purchase behaviour. With high precision, recall, and F1-score values, the model effectively identifies both customers who made a purchase and those who did not. This indicates its potential to assist the retail company in targeting potential buyers more effectively and optimizing its marketing strategies.