📊 Results Summary

✅ Data-Cleaning Decisions

- **Missing Values**: The dataset used "?" to represent missing values. All such rows were removed to maintain data integrity — particularly affecting columns like workclass, occupation, and native_country.
- **Whitespace & Inconsistencies**: Stripped leading/trailing whitespaces from categorical values to avoid redundant category entries (e.g., " Private" vs. "Private").
- **Target Encoding**: The income column was label-encoded into binary form: <=50K as 0 and >50K as 1.
- **Feature Engineering**:
    - Created education_hours by multiplying education_num with hours_per_week, capturing an interaction effect between education level and working time.
    - Created age_group by binning the age column into categorical ranges (e.g., 18–25, 26–35, etc.) to help the model recognize age segments.

📈 Final Model Performance Metrics (Random Forest Classifier)

Metric	Value
Accuracy	0.914
ROC AUC Score	0.967
F1 Score	0.910
Precision (Class 1)	0.95
Recall (Class 1)	0.87
Support (Class 1)	4,824
Precision (Class 0)	0.88
Recall (Class 0)	0.95
Support (Class 0)	4,898

Confusion Matrix:

	Predicted: 0	**Predicted: 1**
Actual: 0	4,677	221
Actual: 1	612	4,212

 Insight: The model performs well across both classes, with a slightly higher false negative rate (612 cases where income >50K was predicted as <=50K). However, the **high precision for high-income predictions (0.95)** makes it particularly

valuable for applications like financial targeting, taxation audits, or premium service segmentation.

🔍 Insights on Top 5 Feature Importance (Random Forest Classifier)

1. **Age** *(Importance: 0.1519)*
   Age is the strongest predictor of income. It likely reflects the typical relationship between professional experience and earning potential.
2. **fnlwgt** *(Importance: 0.1510)*
   While not directly related to income, this census weight feature may indirectly encode socio-demographic trends influencing income, such as location-based or population group effects.
3. **Capital Gain** *(Importance: 0.0772)*
   Capital gains significantly correlate with higher income, suggesting that individuals earning through investments are more likely to surpass the $50K threshold.
4. **Marital Status: Married-civ-spouse** *(Importance: 0.0629)*
   Married individuals, particularly in civil unions, are often more financially stable or further along in their careers — both indicators of higher income.
5. **Hours per Week** *(Importance: 0.0432)*
   Consistent with expectations, longer working hours are generally associated with higher income, though this may vary depending on job type and compensation model.

💡 **Overall Insight**: The Random Forest model captures both intuitive and indirect drivers of income — from age and hours worked to socio-demographic factors — resulting in **robust and accurate predictions**.