Class 5 Assignment

Siravich Pipitarungsri
6758105056

1. Data Preparation

The tourism dataset contains 4,128 customer records with 21 features. The target variable is **ProdTaken**, representing whether a customer purchased the tourism product (1) or not (0).

Manual Data Cleaning (Excel)

Before implementing the machine learning pipeline, the dataset was manually cleaned using Microsoft Excel to inspect structural issues. The cleaning process included:

- Removing duplicate rows
- Removing empty rows
- Verifying column headers
- Checking categorical consistency
- Saving the cleaned dataset as `tourism_clean.csv`

Programmatic Validation (Python)

After importing the cleaned dataset into Python using pandas, I performed validation checks:

- Dataset shape: (4128, 21)
- Missing values: 0
- Duplicate rows: 0

Class distribution:

- Class 0 (Non-buyer): 3,331 samples (80.69%)

- Class 1 (Buyer): 797 samples (19.31%)

The dataset is imbalanced, with buyers representing only 19.31% of total samples.

To preserve class proportions, a **stratified train-test split (75% training / 25% testing)** was applied.

results > dataset_overview.txt

```
Data shape: (4128, 21)
Total missing values: 0
Duplicate rows: 0

Class distribution (ProdTaken):
ProdTaken
0    3331
1     797

Class distribution (%):
ProdTaken
0    80.69
1    19.31

```

(File: `results/dataset_overview.txt`)

Caption: Figure 1. Dataset overview and class distribution after cleaning.

2. Analysis

Exploratory inspection reveals:

- The dataset contains demographic and behavioral features.
- The target variable is highly imbalanced.
- A naive classifier predicting all customers as non-buyers would achieve approximately 80.69% accuracy but 0% recall for buyers.

Therefore, accuracy alone is not sufficient.
Evaluation must prioritize **recall and F1-score for the minority class (buyers)**.

This observation influenced model selection and threshold optimization strategy.

3. Feature Extraction

The dataset contains both categorical and numerical features.

Categorical Features

Applied **One-Hot Encoding** using `OneHotEncoder(handle_unknown="ignore")`:

- Converts categorical variables into binary vectors
- Prevents ordinal bias
- Allows the model to process categorical data numerically

Numerical Features

Applied **StandardScaler**:

- Normalizes feature scales
- Improves model convergence
- Ensures stable optimization

All features were retained for modeling. No manual feature selection was applied.

4. Building Model

Two models were evaluated:

- Logistic Regression (baseline)
- Random Forest (final selected model)

The final selected model is:

RandomForestClassifier

Configuration:

- n_estimators = 400
- class_weight = "balanced_subsample"
- random_state = 42

Because the dataset is imbalanced, class weighting was applied.

Additionally, probability threshold tuning was performed.

Instead of using the default classification threshold (0.5), a validation split was used to search for the threshold that maximizes F1-score for the buyer class.

The optimal threshold found was:

Threshold = 0.22

This significantly improved minority class performance compared to the baseline model.

5. Evaluation Results

Final test performance:

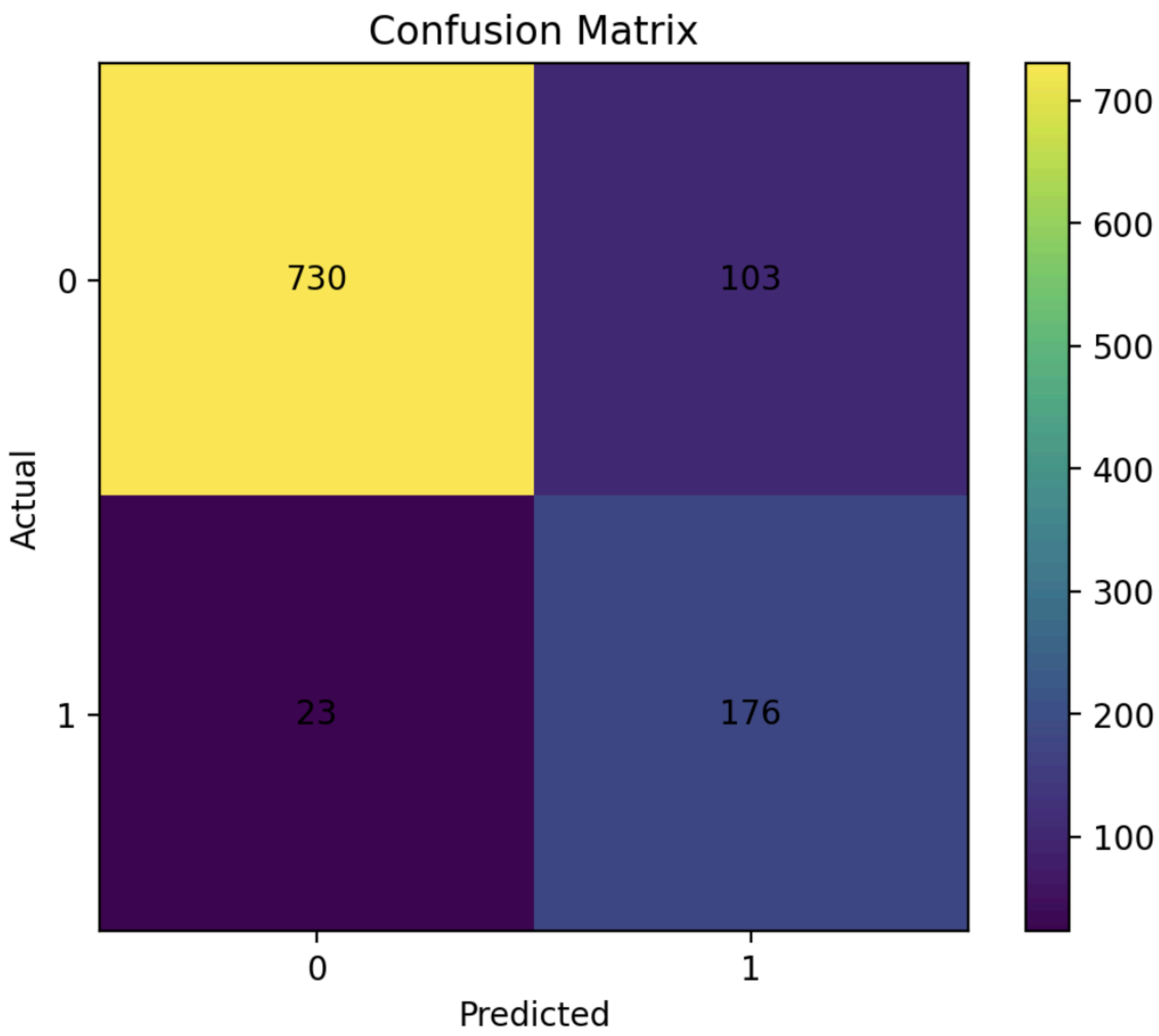
- Accuracy: 87.79%
- Precision (Buyer): 0.63
- Recall (Buyer): 0.88

- F1-score (Buyer): 0.736

Confusion Matrix (Counts):

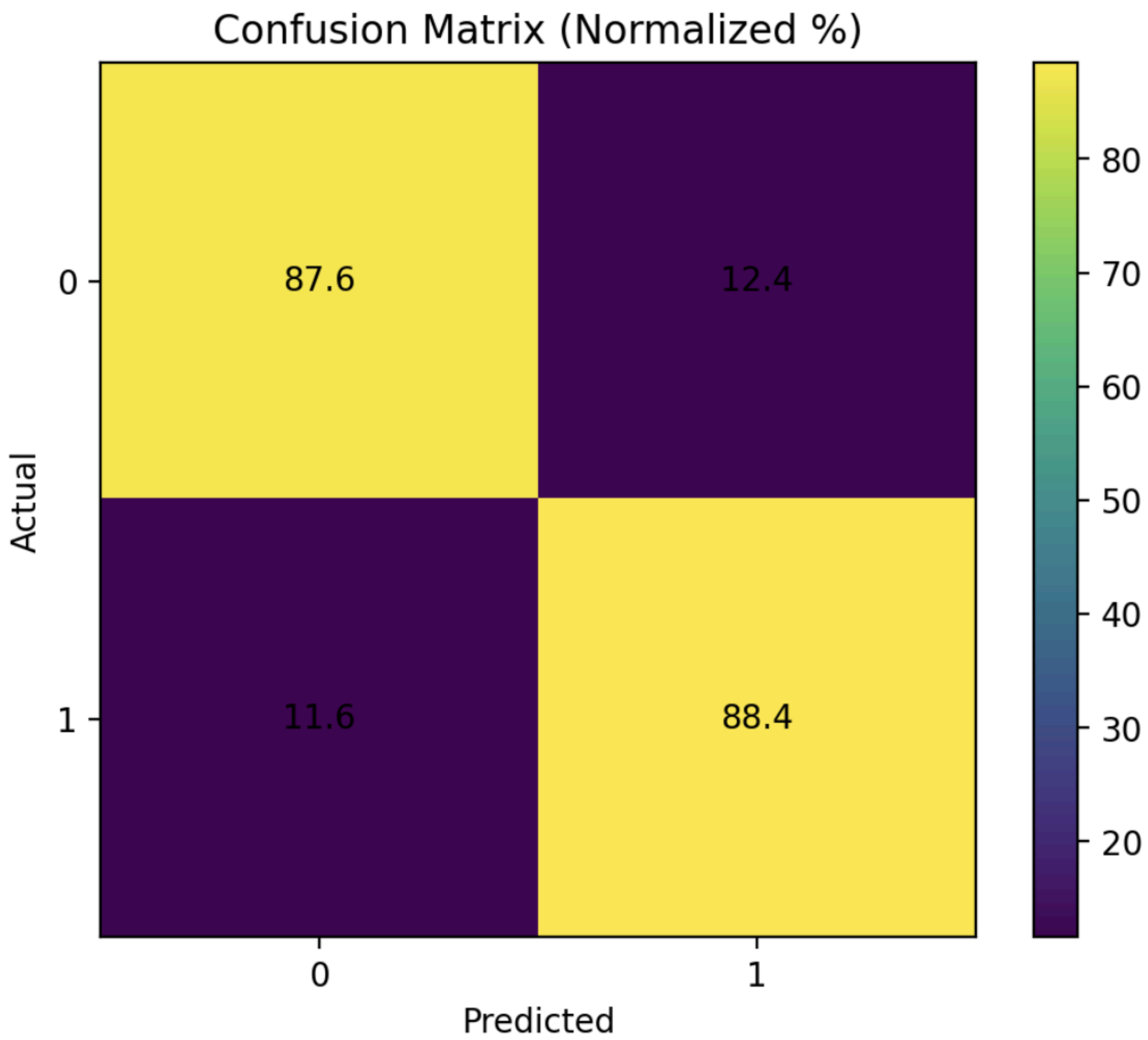
- True Negatives (TN): 730
- False Positives (FP): 103
- False Negatives (FN): 23
- True Positives (TP): 176

The model correctly identifies 176 buyers and misses only 23 buyers.



(File: `results/confusion_matrix.png`)

Caption: Figure 2. Confusion matrix (test set results).



(File: `results/confusion_matrix_normalized.png`)

Caption: Figure 3. Normalized confusion matrix showing class-wise recall.

The normalized confusion matrix shows:

- Recall for Class 0: 87.6%
- Recall for Class 1 (Buyer): 88.4%

Compared to a naive baseline model (80.69% accuracy but 0% buyer recall), the final model provides meaningful predictive power.

```
results > tourism_evaluation.txt
Best Model: RandomForest
Threshold: 0.22

Accuracy: 0.877907
Precision (1): 0.630824
Recall (1): 0.884422
F1-score (1): 0.736402

Confusion Matrix:
[[730 103]
 [ 23 176]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.88      0.92       833
           1       0.63      0.88      0.74       199

    accuracy                           0.88      1032
   macro avg       0.80      0.88      0.83      1032
weighted avg       0.90      0.88      0.89      1032
```

(File: Screenshot of `results/tourism_evaluation.txt`, crop evaluation section only)

Caption: Figure 4. Final classification report for Random Forest with threshold tuning.

The classification report confirms:

- Strong recall for buyers (88%)
- Balanced precision (63%)
- Competitive F1-score (0.736)

Although precision is moderate, recall is prioritized because in marketing applications missing potential buyers results in lost revenue opportunities. Therefore, this trade-off is strategically appropriate.

Conclusion

A complete machine learning pipeline was implemented:

- Manual cleaning in Excel
- Programmatic validation in Python
- Feature encoding and scaling
- Model comparison
- Threshold optimization
- Comprehensive evaluation

The final Random Forest model achieved:

- 87.79% overall accuracy
- 88% recall for buyers
- 0.736 F1-score for buyers

The pipeline is structured, reproducible, evidence-based, and supported by quantitative evaluation results.

The final submitted implementation is located in:

`src/assignment5_tourism_f1.py`