VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Probability and Statistics

Final Project Report

Internet Advertisement Classification using Random Forest

Advisor(s): [Instructor Name]
Student(s): [Student Name 1] [Student ID 1]
[Student Name 2] [Student ID 2]
[Student Name 3] [Student ID 3]

HO CHI MINH CITY, AUGUST 2025

Contents

1 Introduction **4**

2 Data Description **4**
2.1 Dataset Characteristics 4

3 Data Cleaning **4**

4 Descriptive Statistics **5**
4.1 Target Variable Distribution 5
4.2 Statistical Plots 5

5 Objective and Methodology **5**
5.1 Objective 5
5.2 Statistical Methods Used 6

6 Main Results **6**
6.1 Model Performance 6
6.2 Random Forest Confusion Matrix 6
6.3 Feature Importance 7
6.4 Results Analysis 7

7 Conclusion **7**
7.1 Member Contributions 7

List of Figures

List of Tables

6.1 Model performance comparison 6

Listings

3.1 R code for data cleaning 5
6.1 R code for Random Forest 6

1 Introduction

Internet advertisement classification is a crucial problem in the field of data processing and machine learning. With the rapid development of the Internet, automatic recognition and classification of advertisements helps improve user experience and optimize content display.

This report presents the analysis process of the "Internet Advertisements Data Set" from the UCI Machine Learning Repository[4] to build a classification model aimed at predicting whether an image is an advertisement or not.

2 Data Description

The "Internet Advertisements" dataset consists of 3279 samples with 1559 columns. Each row in the data represents an image labeled as "ad" (advertisement) or "nonad" (non-advertisement) in the last column. Columns 0 to 1557 represent numerical attributes of the image.

2.1 Dataset Characteristics

- Number of samples: 3279
- Number of features: 1558
- Target variable: binary (ad/nonad)
- Missing values: represented by "?"

3 Data Cleaning

The data cleaning process includes the following steps:

1. **Handling missing values**: Replace "?" values with the median of each column
2. **Standardizing target variable**: Ensure class labels have consistent format
3. **Removing index columns**: Delete unnecessary index columns for classification

Listing 3.1: R code for data cleaning

```r
# Handle missing values
for(i in 1:(ncol(data)-1)) {
  if(any(data[[i]] == "?")) {
    data[[i]] <- as.numeric(ifelse(data[[i]] == "?", NA, data
       [[i]]))
    data[[i]][is.na(data[[i]])] <- median(data[[i]], na.rm =
       TRUE)
  }
}
```

4 Descriptive Statistics

Descriptive statistical analysis was performed to better understand the data distribution and relationships between variables.

4.1 Target Variable Distribution

The dataset has class imbalance with approximately 1:6 ratio between "ad" and "nonad" classes.

4.2 Statistical Plots

- **Histogram**: Display feature distributions
- **Boxplot**: Detect outliers
- **Scatter plot**: Explore relationships between variables
- **Correlation heatmap**: Display correlation matrix

5 Objective and Methodology

5.1 Objective

Build a classification model to accurately predict whether an image is an advertisement or not based on numerical features.

5.2 Statistical Methods Used

The project uses three main machine learning methods:

1. **k-Nearest Neighbors (k-NN)** [2]: Simple distance-based method
2. **Decision Tree** [5]: Interpretable model with tree structure
3. **Random Forest** [1]: Ensemble method combining multiple decision trees

Reasons for choosing Random Forest as the main method:

- Handles high-dimensional data well
- Reduces overfitting compared to single Decision Trees
- Provides feature importance information
- High performance on various types of data

6 Main Results

6.1 Model Performance

Table 6.1: Model performance comparison

Model	Accuracy	Precision	Recall
k-NN (k=3)	0.81	0.99	0.75
Decision Tree	0.91	0.89	0.47
Random Forest	0.91	1.00	0.38

6.2 Random Forest Confusion Matrix

Listing 6.1: R code for Random Forest

```r
# Build Random Forest with 100 trees
rf_model <- randomForest(x = X_train, y = y_train,
                         ntree = 100,
                         mtry = floor(sqrt(ncol(X_train))))
```

```
5
6 # Predict on test set
7 y_pred <- predict(rf_model, X_test)
```

6.3 Feature Importance

Random Forest shows the most important features in advertisement classification, with X2, X1243, and X1 being the three features with the greatest impact.

6.4 Results Analysis

- **Random Forest** achieved the highest accuracy (91%) with perfect precision (100%)
- **Decision Tree** provides good balance between performance and interpretability
- **k-NN** shows stable performance with simple approach
- All models handle the classification problem well with accuracy above 80%

7 Conclusion

The project has successfully applied machine learning methods to classify Internet advertisements. Random Forest was identified as the best method with 91% accuracy and perfect precision, suitable for deployment in real advertisement filtering systems.

Using Random Forest - a method not previously learned in class - has provided deep insights into handling high-dimensional data and the importance of combining multiple models to improve performance.

7.1 Member Contributions

Member 1 : Data preprocessing and descriptive statistical analysis

Member 2 : Building and evaluating machine learning models

Member 3 : Report writing and result presentation

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[4] M. Lichman. UCI machine learning repository, 2013.

[5] J Ross Quinlan. *C4. 5: programs for machine learning*. Morgan kaufmann, 1993.