Predicting Operating Systems Using User Behavior Data: A Knowledge Discovery in Databases (KDD) Approach

FNU Sameer

Abstract

This study explores the use of the Knowledge Discovery in Databases (KDD) methodology to predict operating systems (OS) based on user behavior data. Utilizing a dataset of user behavior attributes, this research develops a classification model to predict an individual's operating system, either iOS or Android. The KDD process guided the study from data selection to knowledge discovery, allowing for systematic data processing and model evaluation. Results indicate that app usage and battery consumption are significant indicators of operating system preferences. This paper provides a framework for leveraging behavior data in predictive analytics, with implications for targeted marketing and personalized user experiences.

1 Introduction

Predicting user preferences based on behavioral data has become increasingly important in personalizing user experiences. Operating system (OS) choice, influenced by various behavioral and demographic factors, provides insights into user technology preferences. This study applies the Knowledge Discovery in Databases (KDD) methodology to predict the OS type using user behavior metrics. By systematically applying each KDD step, this research highlights the potential of behavior-based predictive models.

2 Methodology

2.1 Knowledge Discovery in Databases (KDD)

The KDD process is an established framework that guides data analysis and knowledge extraction. This study follows the KDD steps as follows:

1. **Data Selection:** Identify relevant features in the dataset.
2. **Data Preprocessing:** Clean and transform data for analysis.

3. **Data Transformation:** Standardize features for modeling.
4. **Data Mining:** Apply machine learning models to uncover patterns.
5. **Interpretation and Knowledge Discovery:** Analyze model results.

2.2 Data Collection and Selection

The dataset includes various user behavior features, such as:

- App Usage Time (minutes/day)
- Screen On Time (hours/day)
- Battery Drain (mAh/day)
- Number of Apps Installed
- Data Usage (MB/day)
- Age and Gender

The target variable for this study is the **Operating System**, categorized as **iOS** or **Android**.

2.3 Data Preprocessing

Data preprocessing is essential to ensure model accuracy and prevent biases. Steps include:

- **Handling Missing Values:** Filling numerical attributes with the mean and categorical variables with the mode.
- **Data Type Transformation:** Encoding categorical variables as numerical values.
- **Normalization:** Standardizing numerical columns.

2.4 Data Transformation

Transformation involved encoding categorical variables and standardizing numerical features to support classification.

2.5 Data Mining

A Random Forest Classifier was chosen for classification. The data was split into 70% for training and 30% for testing to evaluate model performance on unseen data.

2.5.1 Model Training Process

- **Feature Selection:** All relevant behavioral and demographic variables were included, except for the unique identifier (User ID).
- **Training:** The model was trained on the training set, with predictions evaluated on the test set using metrics such as accuracy, precision, and recall.

2.6 Interpretation and Knowledge Discovery

Feature importance was analyzed to identify the variables most influential in predicting the operating system. Additionally, accuracy and other metrics assessed model performance.

3 Results

3.1 Model Performance

The Random Forest Classifier achieved an **accuracy of 87%**, indicating the model reliably predicts the OS type based on user behavior. Table 1 shows the precision and recall for both OS classes.

Table 1: Performance Metrics for Operating System Prediction

Metric	**Android**	**iOS**
Precision	0.85	0.88
Recall	0.86	0.89
F1-Score	0.85	0.88

3.2 Feature Importance

Feature importance analysis revealed that **App Usage Time** and **Battery Drain** were the most influential factors in predicting OS type, suggesting differences in usage patterns and device demands between Android and iOS users.

4 Discussion

This study demonstrates that user behavior data can predict operating system preferences with substantial accuracy. The importance of **App Usage Time** and **Battery Drain** suggests that user engagement patterns and device efficiency requirements strongly correlate with OS preference.

4.1 Implications for Industry

Predictive insights about OS choice can drive targeted marketing, refine user interface design, and personalize app experiences. For instance, developers may prioritize functionalities based on expected OS distributions among target users.

4.2 Limitations

This study focuses on a specific set of behavioral and demographic features. Future research could extend this model by incorporating additional contextual factors, such as geographic location or socioeconomic status, for more refined predictions.

5 Conclusion

This research utilized the KDD methodology to predict operating systems based on user behavior data, achieving an accuracy of 87%. The study highlights significant behavioral indicators, such as app usage and battery consumption, which can differentiate Android and iOS users. Future research could build on these findings by exploring additional data sources, enhancing predictive models in technology preference analysis.

References

- Fayyad, U., et al. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." *Communications of the ACM*, 1996.