# Datasheet for Ad Click Prediction Dataset

## Motivation

* **Purpose:** This dataset was created for beginners in machine learning and data science to practice binary classification tasks. The objective is to predict whether a customer clicks on an advertisement based on their demographic information.
* **Creators:** The dataset was uploaded to Kaggle by Jahanvee Narang.
* **Entity & Funding:** No specific organization or funding body is credited. It appears to be independently shared for educational purposes.

### ****Composition****

* **Instances Represented:** Each row in the dataset represents a unique user, capturing demographic and behavioral features such as age, gender, device type, ad position, and browsing context, along with a binary target indicating whether the user clicked on an ad.
* **Number of Instances:** The raw dataset contains 10,000 entries and 9 columns before cleaning.
* **Missing Data:** Several columns exhibit missing data — age (~48% complete), gender (~53% complete), device\_type, ad\_position, and time\_of\_day (~80% complete each), while browsing\_history is ~52% complete.
* **Confidential Data:** The dataset includes personally identifiable information such as full\_name and id, as well as user browsing behavior, which may be considered sensitive. These columns should be handled with caution or dropped entirely in modeling pipelines for privacy compliance.

## Collection Process

* **Acquisition Method:** The data appears to have been synthetically generated or anonymized and shared on Kaggle.
* **Sampling Strategy:** Not explicitly stated.
* **Time Frame:** No specific dates are mentioned. The dataset was last updated approximately 4 years ago on Kaggle.

### Preprocessing/Cleaning/Labeling

* **Preprocessing**:
* Irrelevant columns such as 'User ID' and 'full\_name' were dropped to remove noise and potential privacy concerns.
* Feature engineering was applied to enrich the dataset, including binning the 'age' column and creating a composite feature from 'device\_type' and 'time\_of\_day'.
* Categorical variables were cast to category dtype to optimize memory usage and enable better handling in LightGBM.
* Missing values were imputed using domain-appropriate strategies: median for numerical variables (e.g., 'age') and a placeholder category such as 'unknown' for categorical variables (e.g., 'device\_type', 'gender').
* **Model Tuning Note**: After preprocessing, **Bayesian Optimization** was used to efficiently search the hyperparameter space of the LightGBM model. This helped identify optimal configurations that improved performance over manual tuning and class-weight-only models.

## Uses

* **Alternative Uses:**
* The dataset can be used for testing classification algorithms, practicing feature engineering, and comparing model performance under class imbalance (e.g., using SMOTE or class weighting).
* **Risks or Harms:**
* As the dataset is synthetic or anonymized, it minimizes privacy risks. However, demographic-based prediction may carry representational biases, especially if applied beyond the toy use-case context.
* **Misuse:**
* This dataset should not be used to train production-level ad targeting systems or real-world decision-making systems without further validation and external data sources.

## Distribution

**Current Distribution:** Publicly distributed via Kaggle (<https://www.kaggle.com/datasets/jahanvee/ad-click-prediction>).

**License and IP:** The dataset license is not explicitly stated; assumed to be for educational use only.

## Maintenance

* **Maintainer:** The Kaggle contributor (Jahanvee Narang) appears to be the original maintainer. However, no future updates or maintenance are guaranteed.