**Sumara kanwal**

**MS-DS\_131(weekend )**

**Detailed Project Report: Mobile Device Usage and User Behaviour Analysis**

**1. Problem Definition and Objectives**

**Problem Definition:**  
The project aims to analyze mobile device usage patterns and classify user behavior into one of five predefined classes (1 to 5) using a dataset containing various metrics such as app usage time, screen-on time, battery drain, and data consumption.

**Objectives:**

* Predict or classify mobile user behavior into five distinct categories based on usage patterns.
* Identify significant factors influencing user behavior classes.
* Explore relationships between app usage, screen-on time, and other metrics.
* Provide actionable insights for developers to optimize user experiences and improve mobile resource efficiency.

**2. Questions Posed and Their Relevance**

1. **What are the most significant factors influencing user behavior class?**  
   Understanding key drivers helps tailor mobile applications and improve device performance.
2. **Are there any differences in usage patterns between male and female users?**  
   This insight can help developers personalize user experiences
3. **How do app usage time and screen-on time correlate with battery drain?**  
   Exploring this can aid in optimizing energy usage and prolonging battery life.
4. **Can user behavior be accurately classified using this dataset?**  
   Ensures the viability of using machine learning for accurate classification.
5. **Are certain age groups more prone to extreme usage behavior?**  
   Helps target specific demographic groups with tailored solutions.

**3. Data Source and Description**

**Dataset Source:**  
The dataset was obtained from Kaggle, titled *Mobile Device Usage and User Behaviour Dataset*. It consists of 700 samples, each containing metrics related to mobile device usage.

**Key Features:**

* **User ID:** Unique identifier for each user.
* **Device Model:** Model of the smartphone.
* **Operating System:** Device OS (iOS/Android).
* **App Usage Time:** Daily time spent on applications (minutes).
* **Screen-On Time:** Average daily screen-on time (hours).
* **Battery Drain:** Daily battery consumption (mAh).
* **Number of Apps Installed:** Total apps installed on the device.
* **Data Usage:** Daily mobile data consumption (MB).
* **Age:** User's age.
* **Gender:** Gender of the user (Male/Female).
* **User Behaviour Class:** Target variable representing behavior class (1–5).

**4. Data Wrangling and Cleaning Steps**

**Steps Taken:**

1. **Missing Value Handling:**
   * Confirmed no null values in the dataset using df.isnull().sum(). None was found , no need for dropping them or filling with means was required for numerical columns.
2. **Duplicate Removal:**
   * Checked duplicate entries using df.duplicated(). As none was found .
3. **Data Type Conversion:**
   * Ensured numerical and categorical features had correct data types.
4. **Outlier Detection and Removal:**
   * Applied the IQR method to identify and remove outliers for numeric columns while excluding specific variables such as gym and lift.
5. **Standardization:**
   * Normalized numerical features to ensure consistent scaling for machine learning models.
6. **Feature engineering**

**5. Insights from Exploratory Data Analysis (EDA)**

**Key Findings:**

1. **Correlation Analysis:**
   * Strong correlations found between app usage time and screen-on time.
   * Battery drain significantly correlated with screen-on time and app usage .
   * High correlation between data usage and the number of apps installed. As the number of apps installed increases, so does the amount of data usage. This relationship implies that users with more apps on their devices may engage in more online activities, which leads to higher data consumption
   * User Behavior Based on Device Type: Samsung Galaxy S21 users are predominantly male, while female users are more inclined to use the Xiaomi Mi 11. iPhone 12 users tend to be balanced between males and females
   * A positive correlation between data usage and the number of installed apps, showing that the more apps installed on a device, the more data is used
2. **Demographic Analysis:**
   * Males had slightly higher screen-on times compared to females, but app usage time was relatively consistent across genders.
3. **Behavior Class Analysis:**
   * Extreme usage behavior (Class 5) associated with higher data consumption and app usage time.
4. **Visualizations Used:**
   * Correlation heatmap to identify relationships between features.
   * Box plots and IQR to detect outliers and analyze class-wise distributions.
   * Scatter plots to explore feature relationships.

**6. Model Choice, Performance Metrics, and Interpretation**

**Model Choice:**  
Selected machine learning models for classification:

* Logistic Regression
* Random Forest Classifier
* Support Vector Machine (SVM)
* Decision Tree
* Kneighbors

**Model Pipeline:**

1. Split the data into training (70%) and testing (30%) sets.
2. Performed feature selection based on correlation analysis.
3. Applied applied hyperparameter tuning to optimize the model parameters and improve performance by n\_estimators=100, max\_depth=10, and learning\_rate=0.01 optimal values.
4. **Performance Metrics:**

* **Accuracy:** Overall model performance.
* **Precision & Recall:** Evaluated for each class to ensure balanced classification.
* **F1-Score:** Measured the harmonic mean of precision and recall.

**Best Performing Model:**

The **Random Forest**, **Logistic Regression**, **SVM**, and **Decision Tree** models achieved perfect scores across all metrics, with precision, recall, and F1-score all equal to 1.000. This indicates that these models are highly accurate, correctly classifying all positive and negative instances in the dataset without any false positives or false negatives.

In contrast, the **K-Nearest Neighbors (KNN)** model performed slightly lower, with a precision of 0.954, recall of 0.954, and an F1-score of 0.953. While still a strong performer, these scores suggest that KNN had a few misclassifications compared to the other models.

**7. Conclusion and Recommendations**

**Conclusion:**

* App usage time, screen-on time, and battery drain were the most significant factors influencing user behavior.
* Behavioral classification is feasible with high accuracy using the dataset.
* Insights suggest targeted optimization for extreme user classes (e.g., power-saving features for heavy users).

**Recommendations:**

1. Developers should focus on optimizing high-drain applications for heavy users.
2. Marketing strategies can target specific age groups prone to extreme usage.
3. Future studies should include more granular features like app categories.