Student(s): Long Le 2203558

Title of the work: Practical Work 1 – Mobile Device Usage and User

**Behaviour Dataset** 

**Predicting Battery Drain Using Machine Learning** 

**Used algorithms: Random Forest Regressor, Logistic regression** 

# Work description and analysis:

## **Description of the work:**

- The dataset, user\_behavior\_dataset.csv, includes user behaviour and device data.
- The goal is to Predict daily battery drain (in mAh/day) and classify user behaviour into predefined categories based on mobile usage patterns and demographic data.
- The dataset contains features such as app usage time, screen-on time, number of apps installed, data usage, and demographic data. The goal is to develop a predictive model to estimate battery drain and identify the most influential factors contributing to battery consumption.

# Data preparation for the training:

#### **Dataset Overview:**

#### The dataset includes:

- Numerical Features: App Usage Time, Screen On Time, Number of Apps Installed, Data Usage, Age.
- Categorical Features: Device Model, Operating System, Gender.
- Target Variables:
  - Battery Drain (Regression task).
  - User Behavior Class (Classification task).

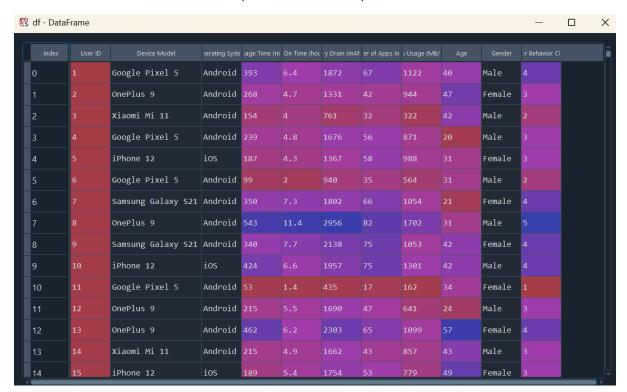


Figure 1. First 14 rows with all columns of training data

# **Preprocessing steps:**

- 1. Dropped Irrelevant Columns: Removed User ID as it has no predictive value.
- 2. Encoded Categorical Variables: Used mapping and OneHotEncoder for categorical features.
- 3. Normalized Numerical Features: Applied StandardScaler for consistent scaling.
- 4. Split Data: Partitioned into training (70%) and testing (30%) sets. Sample Training Data:

Арр	Screen	Number of	Data	Age	Device	Operating	Gender
Usage	On Time	Apps Installed	Usage		Model	System	
Time							
393	6.4	67	1122	40	Google	Android	Male
					Pixel 5		
268	4.7	42	944	47	OnePlus 9	Android	Female
154	4.0	32	322	42	Xiaomi Mi	Android	Male
					11		

# Relevant metrics for the case(s):

# **Regression Metrics:**

- Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.
- Mean Squared Error (MSE): Penalizes larger errors more significantly.
- Root Mean Squared Error (RMSE): Square root of MSE, easier to interpret.
- R-squared (R<sup>2</sup>): Proportion of variance explained by the model.

### **Classification Metrics**

- Accuracy: Overall correctness of predictions.
- Confusion Matrix: Breakdown of true positives, false positives, etc.
- Precision, Recall, F1-Score: Measure of classification quality, especially for imbalanced data.

### Conclusions of the results:

### **Random Forest Regressor (Battery Drain Prediction):**

• Performance Metrics:

 $\circ$  **R**<sup>2</sup>: 0.95

o MAE: 148.96 mAh/day

o **RMSE**: 176.77 mAh/day

```
[15 rows x 11 columns]
Mean Absolute Error (MAE): 148.95990476190477
Mean Squared Error (MSE): 31249.32521238095
Root Mean Squared Error (RMSE): 176.7747866987285
R-squared (R<sup>2</sup>): 0.9516723527485257
Predicted Battery Drain for new user: [1415.63] mAh/day
```

Figure 2. Regression (Random Forest) metrics result

#### Model Analysis:

The Random Forest Regressor achieved outstanding performance, explaining 95% of the variance in battery drain. Its low MAE and RMSE indicate consistent and accurate predictions.

#### Model Usability:

The model is practical for real-world applications, such as optimizing app usage to save battery life

### Logistic Regression (User Behaviour Classification):

#### Performance Metrics:

Accuracy: 73.81%

Confusion Matrix:

 Precision, Recall, F1-Score: Moderate to high for most classes, with notable strength in Class 1 and Class 5.

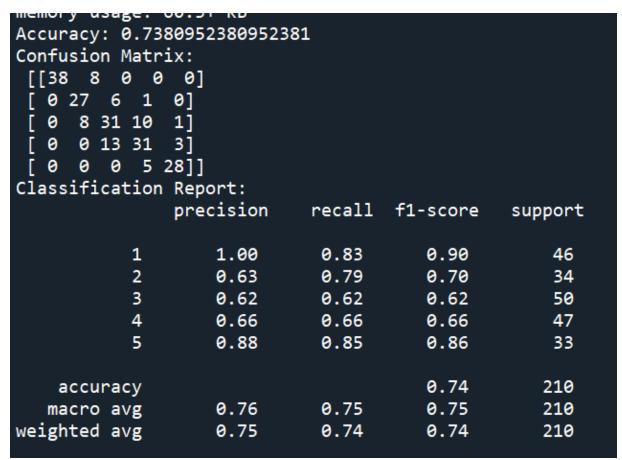


Figure 3. Classification (Logistic Regression) result

## Model Analysis:

Logistic Regression effectively distinguishes certain behaviour classes, such as Class 1 (precision: 1.00) and Class 5 (precision: 0.88). However, moderate performance for Classes 2, 3, and 4 indicates room for improvement.

### **Comparison of Models:**

### • Best Performance:

For regression tasks, the Random Forest Regressor outperformed Linear Regression with an R<sup>2</sup> of 0.95.

For classification, Logistic Regression achieved reasonable accuracy but requires enhancement to reduce misclassifications.

#### Model Usability:

Both models are suitable for practical applications in their respective tasks, with Random Forest providing more reliable predictions.

# **Potential Improvements:**

- 1. Add more diverse and recent data to improve generalization.
- 2. Hyperparameter tuning for Logistic Regression and Random Forest.
- 3. Test advanced algorithms like XGBoost or Gradient Boosting.