MIT | Academy of Engineering

Online Gaming Behavior Analysis Using ML

Understanding how age, gender, engagement, and game difficulty influence playtime, purchases, and achievements.

202301070140 - Vaishnav Lajurkar
202301070141 - Mukund Chavan
202301070142 - Aditya Ganeshe
202301070143 - Prathamesh Shinde

🧩 Problem Statement

With the rise of digital entertainment, online gaming has become one of the fastest-growing industries. However, understanding **player engagement**, **behavioral trends**, and **in-game purchase patterns** is crucial for improving player experience and retention.

This project aims to analyze and predict player engagement levels using machine learning, based on demographic, gameplay, and behavioral data from **40,000 gamers**.

Dataset Used: Online Gaming Behavior Dataset (40,034 records) **Colab Link: Open**

📊 Dataset Description & Preprocessing

Features:

- 🧍‍♂️ Demographics: Age, Gender, Location
- 🎮 Gameplay: Game Genre, Difficulty, Sessions per week, Playtime hours
- 💸 Behavioral: In-game purchases, Engagement level, Achievements unlocked

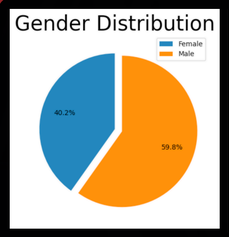
Preprocessing Steps:

- Handled missing and inconsistent data
- Label encoding for categorical attributes (Gender, GameGenre, etc.)
- Normalization using StandardScaler
- Train–test split (80:20) for model evaluation

IoT Relevance (Conceptual):

Smart gaming consoles, sensors, or VR devices could stream real-time engagement data such as session duration, reaction time, or physiological signals — enabling dynamic behavioral prediction.

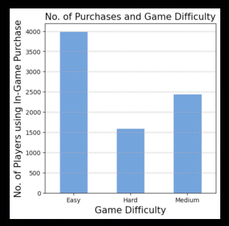
🔍 Key Insights



GameGenre / Location	Action	RPG	Simulation	Sports	Strategy
Asia	1656	1620	1623	1582	1614
Europe	2428	2372	2414	2431	2359
Other	792	787	795	775	786
USA	3163	3173	3151	3260	3253

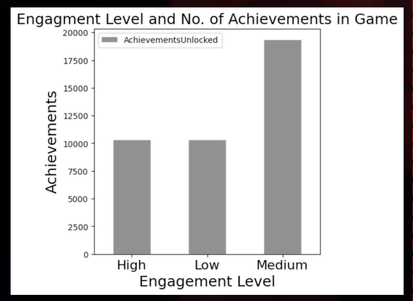
60% M | 40% F Gamers

Regional activity: **USA > Europe > Asia**



Players with easy difficulty and medium engagement make the most purchases.

Medium engagement players unlock the highest achievements.



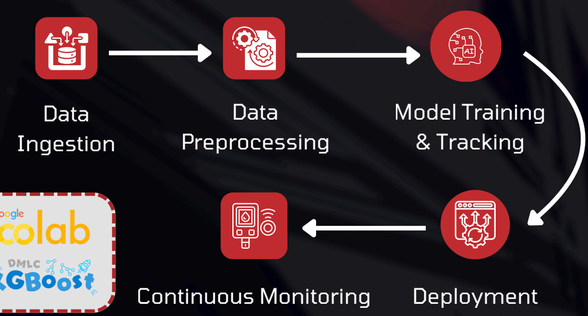
InGamePurchases	mean	std
No	94.87	49.10
Yes	94.49	48.68

Average gaming session duration ≈ 95 minutes/day.

Flowchart / Workflow

Data Collection
↓
Preprocessing
↓
Feature Engineering
↓
Model Training (LR, RF, LGBM, XGB)
↓
Evaluation & Validation
↓
Best Model Selection
↓
Insights & Visualization

MLOps Workflow



Tools & Technologies

pandas, matplotlib, NumPy, Google colab, seaborn, LightGBM, XGBoost

🧠 Model Selection & Comparative Analysis

Model	Type	Strengths	Limitations
Logistic Regression	Baseline	Simple, interpretable	Weak with nonlinear patterns
Random Forest	Ensemble	High accuracy, interpretable features	Computationally heavy
LightGBM	Gradient Boost	Fast, efficient, handles imbalance	Requires tuning
XGBoost	Ensemble Boost	High precision, great generalization	Slower training time

📈 Results & Model Evaluation

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.814	0.815	0.814	0.812
Random Forest	0.877	0.878	0.877	0.877
XGBoost	0.879	0.879	0.879	0.879
LightGBM (Best)	**0.881**	**0.882**	**0.881**	**0.881**

✅ Best Model: LightGBM

- Achieved the highest accuracy (88%)
- Balanced performance across all metrics