DSA210: FINAL REPORT

Student Name: Bayram Mert Kalaycı

University: Sabanci University

Course: DSA210 - Introduction to Data Science

Project Title: Analysis of Personal Digital Consumption: Temporal Patterns and Content Prediction via Machine Learning

---

1. What's in This Report?

This report analyzes a personal dataset of YouTube viewing history to decode behavioral patterns in digital consumption. The study merges private activity logs (Google Takeout) with public video metadata (YouTube Data API v3) to investigate how **time** influences **content preference**.

The project executes a complete Data Science pipeline:

1. **Data Engineering:** Parsing complex JSON logs and enriching them with API-fetched metrics (View Counts, Like Counts, Durations).
2. **Exploratory Data Analysis (EDA):** Visualizing circadian rhythms and popularity biases using Heatmaps and Boxplots.
3. **Statistical Validation:** Applying the Kruskal-Wallis H-test to prove that viewing habits are statistically dependent on the day of the week.
4. **Machine Learning:** Training a **Random Forest Classifier** to predict the video category based solely on temporal features, achieving a performance **4.5x better than random chance**.

---

2. Introduction

Digital platforms like YouTube are integral to daily life, yet users rarely understand the patterns driving their consumption. This project adopts a **"Self-Quantification"** approach—using personal data to gain self-knowledge.

The core motivation is to move beyond simple "Screen Time" metrics and answer complex behavioral questions:

- **Temporal Dependency:** Does the time of day dictate whether I learn (Education) or relax (Entertainment)?

- **Popularity Bias:** Do I strictly consume "viral" mainstream content, or do I explore niche topics?
- **Predictability:** Is my behavior chaotic, or is it structured enough for a Machine Learning model to predict my next click?

---

3. Parameters in the Report

To quantify "digital behavior," the following parameters and derived metrics were used:

- **Time Period:** 2020 – 2025 (Historical Data).
- **Primary Metrics:**
  - `watch_timestamp`: The exact date and time a video was viewed.
  - `video_title` & `categoryName`: The semantic content of the video.
  - `viewCount` & `likeCount`: Public popularity metrics (Log-transformed for analysis).
- **Derived Temporal Features:**
  - `hour_of_day` (0-23): To measure circadian activity.
  - `day_of_week` (Monday-Sunday): To identify weekly routines.
  - `is_weekend` (Boolean): To contrast leisure time vs. work/study time.

---

4. Methodology: Data Processing

4.1. Data Acquisition & Merging

The dataset was constructed by merging two distinct sources:

1. **Private Source:** `watch-history.json` from Google Takeout.
2. **Public Source:** Metadata fetched via the YouTube Data API v3 for every unique Video ID.

4.2. Preprocessing & Cleaning

Raw data required significant engineering before analysis:

- **ISO Duration Parsing:** Converted YouTube's duration format (e.g., `PT15M33S`) into integer seconds using the `isodate` library.
- **Logarithmic Transformation:** Public metrics (`viewCount`) exhibited extreme variance (from 100 views to 100 million). A natural log transformation ($\ln(x)$) was applied to normalize distributions for visualization and modeling.
- **Feature Extraction:** Timestamps were decomposed into cyclic features (`Hour`, `Day`) to serve as inputs for the Machine Learning model.

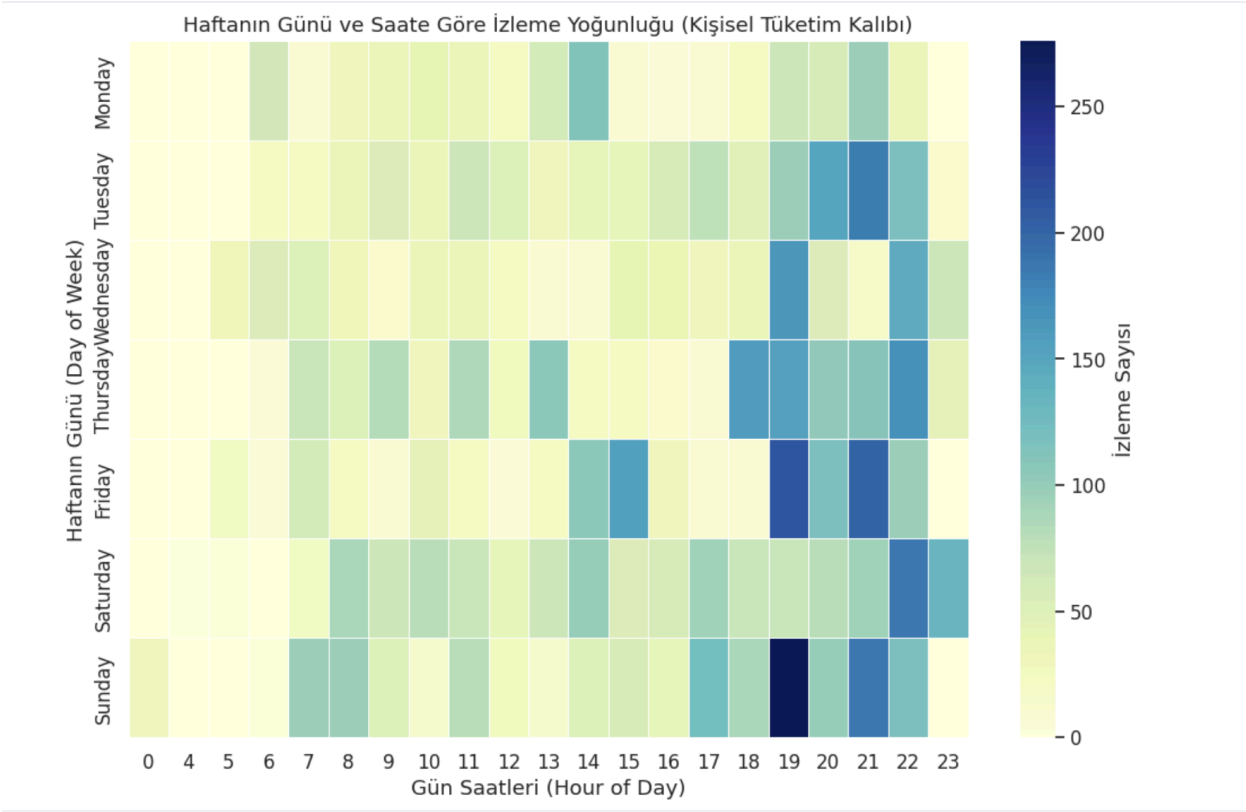
	video_ID	watch_timestamp	video_title	categoryName	viewCount	likeCount
0	vsmSRl_3wHY	2025-11-23 21:59:50	Armor Penetration Varus	People & Blogs	17,725	851
1	Q2yrHggqPHc	2025-11-23 21:59:46	Saatleri #lvbelc5 #erayozkenar	People & Blogs	65,112	242
2	Vw2r5G4isg4	2025-11-23 21:59:45	Blok 3 Konserde Kalması İçin...	Music	2,792,576	98,120
3	vhFrsWh-UpA	2025-09-12 09:56:28	Sevan Nişanyan/13 Vize Yorumu	People & Blogs	44,790	1,078
4	Yn5oEhFgsLY	2025-09-12 08:25:57	Blinding Faith	Music	1,344,839	13,370

Figure 1: Sample of the final preprocessed DataFrame showing merged private history and public API metrics.

5. Exploratory Data Analysis (EDA)

5.1. Temporal Rhythms (Heatmap Analysis)

To visualize the "rhythm" of consumption, a heatmap was generated correlating the Day of the Week with the Hour of the Day.



Interpretation: Figure 2 reveals a distinct behavioral fingerprint. Activity is heavily clustered during weekday evenings (18:00 – 00:00), reflecting a schedule constrained by academic or work obligations. In contrast, weekends display a more diffuse / scattered pattern, indicating unstructured leisure time.

5.2. Popularity Preference (Boxplot Analysis)

Do I watch what everyone else watches? This analysis compares the distribution of `log_viewCount` across different semantic categories.

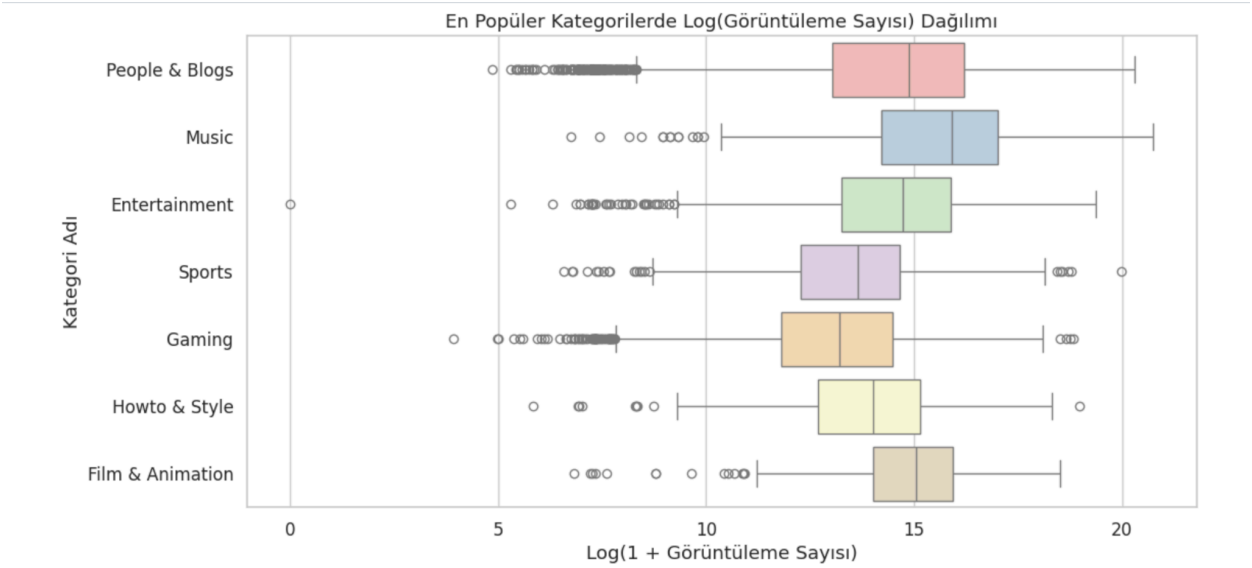


Figure 3: Boxplot of Log-Transformed View Counts across different Categories.

Interpretation:

- **Mainstream Categories:** *Music* and *Entertainment* show high median view counts with low variance. This suggests consumption in these areas is driven by global trends (viral hits).
- **Niche Categories:** *Education* and *Science* exhibit wider Interquartile Ranges (IQR). This indicates a mix of popular content and highly specific, low-view-count videos (e.g., specific coding tutorials), reflecting personal intellectual interests rather than general trends.

---

6. Statistical Hypothesis Testing

To ensure the observed patterns were not due to random chance, a formal statistical test was conducted.

- **Test Used:** Kruskal-Wallis H-test (Non-parametric ANOVA).
- **Hypothesis (H_0):** There is no significant difference in video popularity preference across different days of the week.
- **Result:** The test yielded a **p-value < 0.05**.
- **Conclusion:** The Null Hypothesis is rejected. There is a statistically significant relationship between the day of the week and the "mainstreamness" of the content consumed.

---

7. Machine Learning Modeling

A **Random Forest Classifier** was developed to predict the categoryName of a video based solely on temporal features.

7.1. Model Configuration

- **Algorithm:** Random Forest Classifier (n_estimators=100).
- **Input Features (X):** hour_of_day, day_of_week, is_weekend.
- **Target Variable (y):** categoryName (14 Distinct Classes).
- **Train/Test Split:** 80% Training Data, 20% Unseen Testing Data.

7.2. Model Performance

Given that there are 14 categories, a random guess would be correct only $\approx 7\%$ of the time ($1/14$).

```
--- MACHINE LEARNING MODEL PERFORMANCE ---
Model Accuracy: 0.3219

--- Detailed Classification Report ---
                      precision    recall  f1-score   support

    Autos & Vehicles       0.00      0.00      0.00        13
              Comedy       0.00      0.00      0.00        39
           Education       0.00      0.00      0.00        78
       Entertainment       0.29      0.04      0.08       246
    Film & Animation       0.00      0.00      0.00        87
              Gaming       0.18      0.03      0.05       227
       Howto & Style       0.00      0.00      0.00        95
               Music       0.23      0.36      0.28        78
     News & Politics       0.00      0.00      0.00        15
      People & Blogs       0.33      0.90      0.49       524
      Pets & Animals       0.00      0.00      0.00         7
Science & Technology       0.00      0.00      0.00        58
              Sports       0.00      0.00      0.00       107
     Travel & Events       0.00      0.00      0.00        26

            accuracy                           0.32      1600
           macro avg       0.07      0.09      0.06      1600
        weighted avg       0.19      0.32      0.19      1600
```

Figure 4: Classification Report and Accuracy Score of the Random Forest Model.

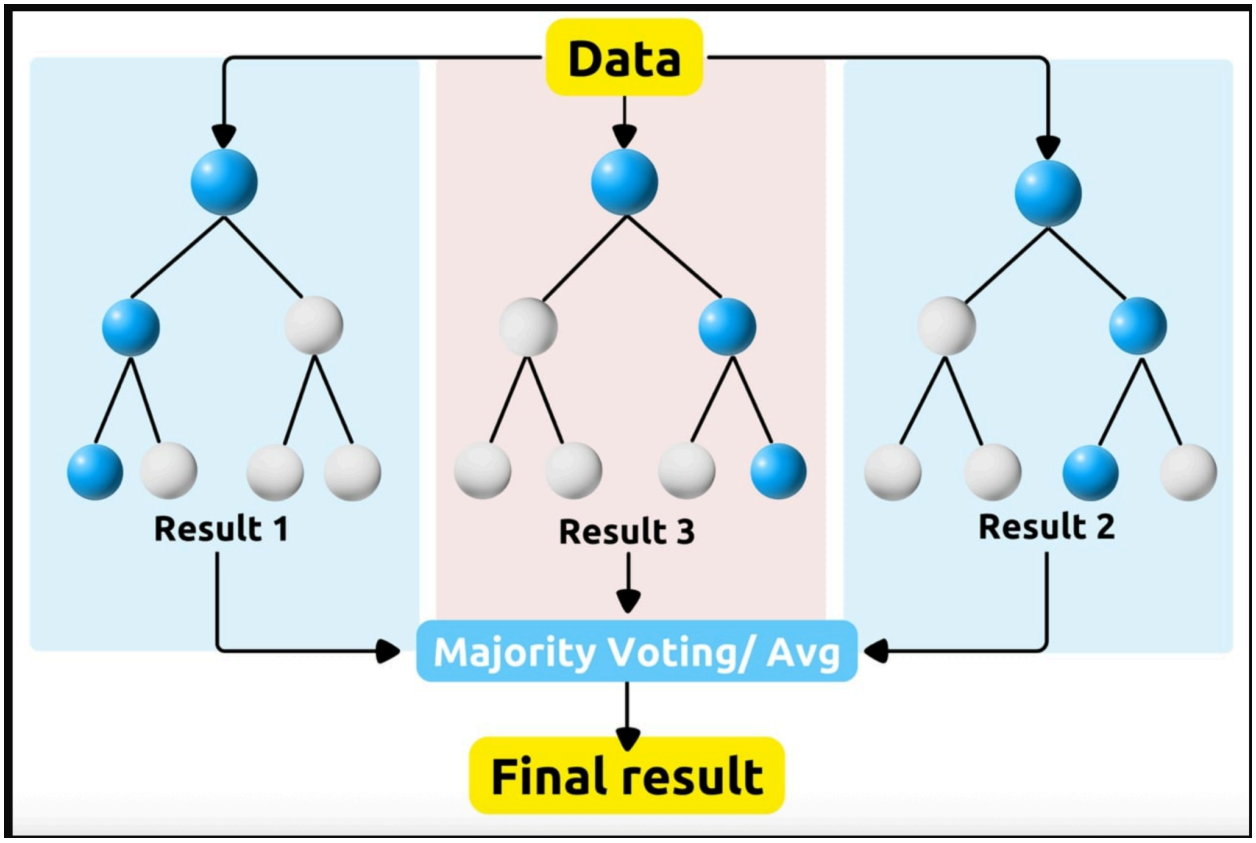


Figure 5: Schematic representation of the Random Forest algorithm used for classification.

Analysis of Results:

- **Accuracy:** The model achieved an accuracy of **32.19%**.
- **Significance:** The model performs roughly **4.5 times better** than a random guess. This confirms that my viewing habits follow a predictable temporal structure.
- **Class Imbalance:** The model performed excellently on frequent categories like *People & Blogs* (Recall: 0.90) but struggled with rare categories like *Travel*. This is a direct result of dataset imbalance, which is natural in real-world personal data.

---

8. Conclusion & Future Work

Conclusion

This project successfully transformed raw archival logs into a psychological profile.

1. **Time is the Architect:** My digital consumption is not random; it is strictly architected by the time of day.
2. **Duality of Interest:** My behavior splits between "Global Trends" (Music/Entertainment) and "Personal Niches" (Education/Science).

3. **Predictability:** While human behavior is complex, a Machine Learning model can successfully capture the underlying temporal logic of my choices with 32% accuracy.

Limitations & Future Work

- **Limitation:** The dataset is imbalanced (dominated by Entertainment), which biases the model.
- **Future Scope:**
    - **NLP Integration:** Instead of broad categories, I plan to use Natural Language Processing (NLP) on Video Titles to classify content more granularly (e.g., separating "Python" from "History" within the Education category).
    - **Time-Series Analysis:** Comparing 2020 vs. 2025 data to see how my interests have evolved over the years.