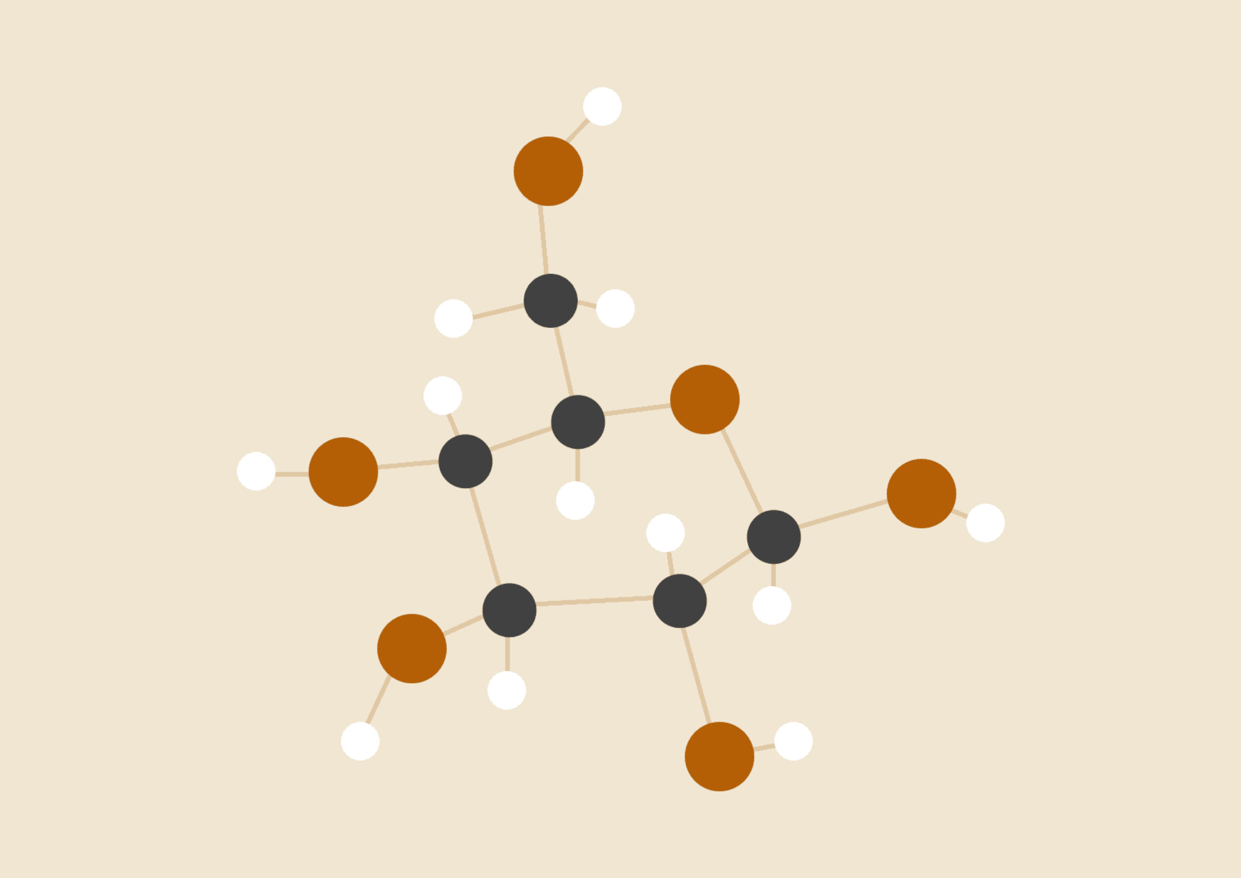
# SPRINT 4 REPORT



**Muhammad Bilal**

30.03.2025

# INTRODUCTION

My dataset is in a CSV-formatted log of **HTTP requests**, likely from a web application (e.g., an **e-commerce site** running on **localhost:8080/tienda1**). The goal is to understand its structure, identify patterns, and spot potential issues for cleaning. Analyzing the columns, data types, unique values, and anomalies.

# DATASET OVERVIEW

My dataset has 17 columns. Each representing a field in an HTTP request log, with a mix of normal and anomalous entries.

## Structure

1. **(Unnamed first column):** Empty in the header, possibly an index or placeholder.
2. **Method:** HTTP method (e.g., GET, POST, PUT).
3. **User-Agent:** Browser or client identifier.
4. **Pragma:** Cache directive (e.g., no-cache).
5. **Cache-Control:** Cache control header (e.g., no-cache).
6. **Accept:** Accepted content types.
7. **Accept-encoding:** Supported encodings (e.g., gzip, deflate).
8. **Accept-charset:** Supported character sets (e.g., utf-8).
9. **language:** Language preference (e.g., en).
10. **host:** Target host (e.g., localhost:8080).
11. **cookie:** Session cookie (e.g., JSESSIONID).
12. **content-type:** MIME type of the request body (e.g., application/x-www-form-urlencoded).
13. **connection:** Connection type (e.g., close).
14. **lenght:** Likely "Content-Length" misspelled; length of the request body.
15. **content:** Request payload (e.g., form data).
16. **classification:** Label (0 for Normal, 1 for Anomalous).
17. **URL:** Full request URL.

## Column-by-Column Exploration

### 1. (Unnamed first column)

1. **Observation:** Empty in the header, but all rows start with a comma, suggesting it's a placeholder or index.
2. **Unique Values:** None (empty).
3. **Issue:** Unclear purpose; likely an artifact of CSV formatting.

### 2. Method

1. **Values:** GET, POST, PUT.
2. **Normal:** Mostly GET and POST.
3. **Anomalous:** Includes PUT (e.g., row with id=1), which is unusual for a typical e-commerce flow.
4. **Insight:** PUT in anomalous data might indicate attempts to modify resources improperly.

### 3. User-Agent

1. **Value:** Uniformly Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko).
2. **Observation:** No variation, suggesting a controlled test environment or a single client.
3. **Issue:** Lack of diversity limits real-world applicability; could be a synthetic dataset.

### 4. Pragma & 5. Cache-Control

1. **Values:** Both are consistently no-cache.
2. **Observation:** No variation; requests disable caching.
3. **Insight:** Expected for dynamic web apps, but uniformity suggests no edge cases.

## 6. Accept

1. **Value:** "text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5".
2. **Observation:** Consistent across all rows.
3. **Insight:** Standard for a browser; no anomalies here.

## 7. Accept-encoding

1. **Value:** "x-gzip, x-deflate, gzip, deflate".
2. **Observation:** Uniform, supports compression.
3. **Insight:** No issues, typical for HTTP requests.

## 8. Accept-charset

1. **Value:** "utf-8, utf-8;q=0.5, *;q=0.5".
2. **Observation:** Consistent, prioritizes UTF-8.
3. **Insight:** Matches language (en), no variability.

## 9. language

1. **Value:** en (English).
2. **Observation:** No variation.
3. **Insight:** Single-language dataset, limits localization testing.

## 10. host

1. **Value:** localhost:8080 (except one PUT request with localhost:9090).
2. **Observation:** Mostly consistent, 9090 in an anomalous PUT request.
3. **Issue:** Port mismatch could indicate a misconfiguration or attack attempt.

## 11. cookie

1. **Value:** JSESSIONID with unique hex values (e.g., 1F767F17239C9B670A39E9B10C3825F4).
2. **Observation:** Each request has a unique session ID.
3. **Insight:** Expected for session tracking; no obvious reuse or anomalies.

## 12. content-type

1. **Values:** Empty for GET, application/x-www-form-urlencoded for POST/PUT.
2. **Observation:** Matches method (empty for GET, present for POST/PUT).
3. **Issue:** One anomalous PUT has an empty Accept header but a content-type, which is inconsistent.

## 13. connection

1. **Values:** close or Connection: close.
2. **Observation:** close for GET, Connection: close for POST/PUT.
3. **Insight:** Minor formatting inconsistency; likely same intent.

## 14. lenght

1. **Observation:** Misspelled "length"; values are integers (e.g., 68, 4) or empty.
2. **Normal:** Empty for GET, matches content length for POST.
3. **Anomalous:** Consistent with POST/PUT content.
4. **Issue:** Spelling error needs correction.

## 15. content

1. **Observation:** Empty for GET; URL-encoded form data for POST/PUT (e.g., id=3&nombre=Vino+Rioja).
2. **Normal:** Typical e-commerce actions (add to cart, login, register).
3. **Anomalous:** Includes suspicious payloads (e.g., <script> tags, malformed parameters like id=1%2F).
4. **Insight:** Anomalous entries suggest injection attempts (XSS, SQLi).

## 16. classification

1. **Values:** 0 (Normal), 1 (Anomalous).
2. **Observation:** Majority are 0; last section is 1.
3. **Insight:** Binary label for anomaly detection; useful for supervised learning.

### 17. URL

1. **Observation:** Mostly http://localhost:8080/tienda1/... with JSP pages or static files.
2. **Normal:** E-commerce actions (e.g., /publico/anadir.jsp, /publico/pagar.jsp).
3. **Anomalous:** Unusual paths (e.g., 6909030637832563290.jsp, .bak files, external redirects).
4. **Insight:** Anomalous URLs suggest probing or exploitation attempts.

## Key Patterns and Insights

1. **Normal Data:** Represents typical e-commerce interactions (browsing, adding to cart, logging in, registering, paying).
2. **Anomalous Data:** Includes potential attacks:
   - **XSS** (e.g., <script>document.location=... in login or modo).
   - **SQL Injection** (e.g., tscha'/j+a in login).
   - **Path Traversal** (e.g., .bak, .Inc).
   - **Invalid Parameters** (e.g., id=1%2F, precioA instead of precio).
   - **Unusual Methods** (PUT) and Ports (9090).
3. **Consistency:** Headers like User-Agent, Accept, etc., are uniform, suggesting a test environment.
4. **Encoding:** URLs and content use percent-encoding (e.g., %F1 for ñ), requiring decoding for analysis.