# Advanced Machine Learning Project Proposal

---

## 1. Team Details

**Team Name:**
Anomaly Archers

**Section:**
B

**Team Members (Exactly 5):**

| S.No | Name | Roll Number | Email ID |
|---|---|---|---|
| 1 | Mohit Kumar | 24BCS10222 | mohit.24bcs10222@sst.scaler.com |
| 2 | Krishna Faujdar | 24BCS10217 | krishna.24bcs10217@sst.scaler.com |
| 3 | Tirth Shah | 24BCS10347 | tirth.24bcs10347@sst.scaler.com |
| 4 | Abhi Gandhi | 24BCS10397 | abhi.24bcs10397@sst.scaler.com |

**Team Leader Name & Email ID:**
Tirth Shah – tirth.24bcs@sst.scaler.com

---

# 2. Project Title

**Predictive Maintenance and Health Categorization for Rental Bike Fleets Using Usage Pattern Analysis**

---

# 3. Problem Statement

## 3.1 Background & Context

Urban bike rental companies operate fleets consisting of thousands of bicycles that are used by diverse users under varying conditions. These bikes experience different levels of wear depending on usage intensity, trip duration, terrain, and frequency.

Currently, most fleet operators rely on **reactive maintenance** (fixing bikes only after failure) or **fixed-interval servicing**, both of which fail to reflect real-world usage patterns. Since maintenance logs are often disconnected from usage data and lack explicit failure labels, identifying bikes that require urgent attention becomes difficult.

This problem is prominent in the **urban mobility and shared transportation industry**, where operational efficiency and customer trust are directly tied to fleet reliability.

---

## 3.2 Core Problem Definition

The core problem is to **identify and prioritize rental bikes that are likely to be mechanically degraded or at risk of failure**, using only historical usage data, without relying on labeled maintenance outcomes.

Specifically, the project aims to:

- Discover distinct **usage-based health profiles** of bikes.
- Detect **abnormal usage behavior** that may indicate hidden mechanical issues.
- Track **temporal degradation trends** in bike performance.

Supervised learning is not suitable because there are **no explicit labels indicating bike health or failure**, making unsupervised learning and anomaly detection the most appropriate approaches.

---

# 4. Dataset Description

## 4.1 Dataset Source

- **Dataset Name:** Capital Bikeshare System Data
- **Source:** Public Open Data Portal (Washington D.C. Bikeshare)
- **Is the dataset public?:** Yes

---

## 4.2 Dataset Overview (Tentative)

| Attribute | Details |
|---|---|
| Number of rows | Millions of trip records |
| Number of columns | ~10–12 |
| Type of data | Time-series, numerical, categorical |
| Time period | Multiple years of historical data |
| Missing values | Minimal (mostly station-related) |

**Key Attributes Include:**

- Trip Duration
- Start & End Date-Time
- Start & End Station
- Bike Number (Unique Asset ID)

---

## 4.3 Why This Dataset Fits the Problem

This dataset represents **real-world operational usage** of rental bikes at scale. While it does not include direct maintenance labels, it allows engineering of meaningful **usage-based health proxies** such as:

- Cumulative ride duration per bike
- Usage frequency per day
- Average trip length and stress indicators

These proxies enable discovery of latent patterns related to mechanical wear and abnormal behavior.

---

# 5. Chosen Advanced ML Technique(s)

### Primary Technique (Mandatory)

☑ **Unsupervised Learning**
☑ **Recommender / Ranking Logic (Maintenance Priority)**
☑ **Time Series Analysis**

---

### Why These Techniques Are Appropriate

- **Unsupervised Learning (K-Means Clustering):**
  Used to segment bikes into natural **usage profiles** such as heavy-commute bikes vs. occasional-use bikes.
- **Anomaly Detection (Isolation Forest):**
  Helps identify bikes whose usage behavior significantly deviates from others in the same cluster, indicating potential mechanical issues.
- **Time Series Analysis:**
  Enables detection of **performance degradation trends** over time by analyzing rolling averages and usage patterns per bike.

Together, these methods allow actionable insights without requiring labeled failure data.

---

# 6. Expected Outcomes & Insights

- **Bike Health Segmentation:**
  Bikes categorized into clusters based on usage intensity and stress.

- **Health Scorecard:**
  Each Bike ID assigned a health category such as **Stable**, **Warning**, or **Critical**.
- **Maintenance Priority Ranking:**
  A ranked list recommending which bikes should be serviced first.
- **Business Impact:**
    - Reduced unexpected breakdowns
    - Improved fleet availability
    - Data-driven maintenance planning

**Success Measurement:**

- Reduction in anomaly concentration over time
- Stability of cluster assignments
- Interpretability and operational usefulness of health scores (not accuracy alone)

---

# 7. Assumptions & Limitations

**Assumptions:**

- Abnormal usage patterns correlate with mechanical degradation.
- Usage intensity is a reasonable proxy for wear.

**Limitations:**

- Lack of actual maintenance logs may introduce false positives.
- External factors such as weather may affect trip behavior.
- Findings are probabilistic, not deterministic diagnoses.

---

# 8. Project Scope Confirmation

☑ We confirm that the dataset is self-sourced
☑ We understand that the problem statement and dataset cannot be changed after approval
☑ We understand that late submissions will not be accepted
☑ Team size is exactly 5 members

---

## 9. Declaration

We confirm that:

- This proposal is our original work
- We will begin project execution only after approval
- We will strictly follow all project guidelines

---