Project Writeup: Personalized Activity/Workout Recommendation System

Team: **ML Mavericks**

Team Members:

- Shubham Kothiya
- Yugm Patel
- Soumya Bharathi Vetukuri

Homework Assignment:

End-to-End Analysis: Changing Data Distribution and Impact on Model Metrics

1. Introduction

This project explores the impact of modifying data distributions on the performance of machine learning models. By up sampling and down sampling, we analyze how it affects classification metrics such as the F1-score. An interactive dashboard is built to dynamically modify the distribution of a selected feature and visualize changes in model performance in real time.

2. Problem Statement

Machine learning models often face challenges due to **imbalanced data distributions**, leading to biased predictions. This project aims to:

- **Modify the distribution of an important feature** using upsampling and downsampling techniques.
- **Train and evaluate models using a Muller Loop approach** to establish baseline performance.
- **Visualize changes in model metrics dynamically** as the feature distribution is modified.
- **Compare the effects of different distributions** on model performance metrics.

3. Dataset Overview

The dataset used is **dailyActivity_merged.csv**, which contains physical activity metrics from users. Key attributes include:

- **TotalSteps** – The total number of steps taken.
- **VeryActiveMinutes** – The time spent in intense physical activity.
- **Calories** – The number of calories burned.
- **ActivityDate** – The date when the data was recorded.

Target Variable Selection

- The dataset does not contain a predefined classification target.
- **Calories are bucketed into three categories** using quantile-based discretization to form a multi-class classification problem.

4. Data Preprocessing & Feature Engineering

Steps Taken:

- Converted ActivityDate to DateTime format.
- Handled missing values (if any).
- Created TargetClass from Calories:
- Low calorie burn → 0
- Medium calorie burn → 1
- High calorie burn → 2
- Identified VeryActiveMinutes as a key feature based on feature importance analysis.

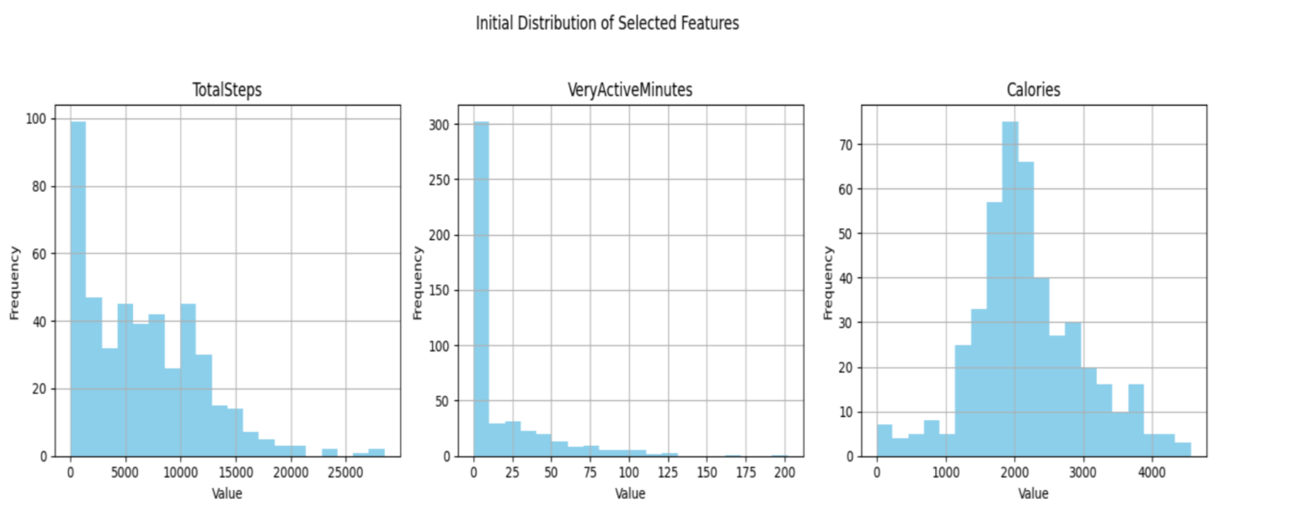
5. Key Features & Feature Importance Analysis

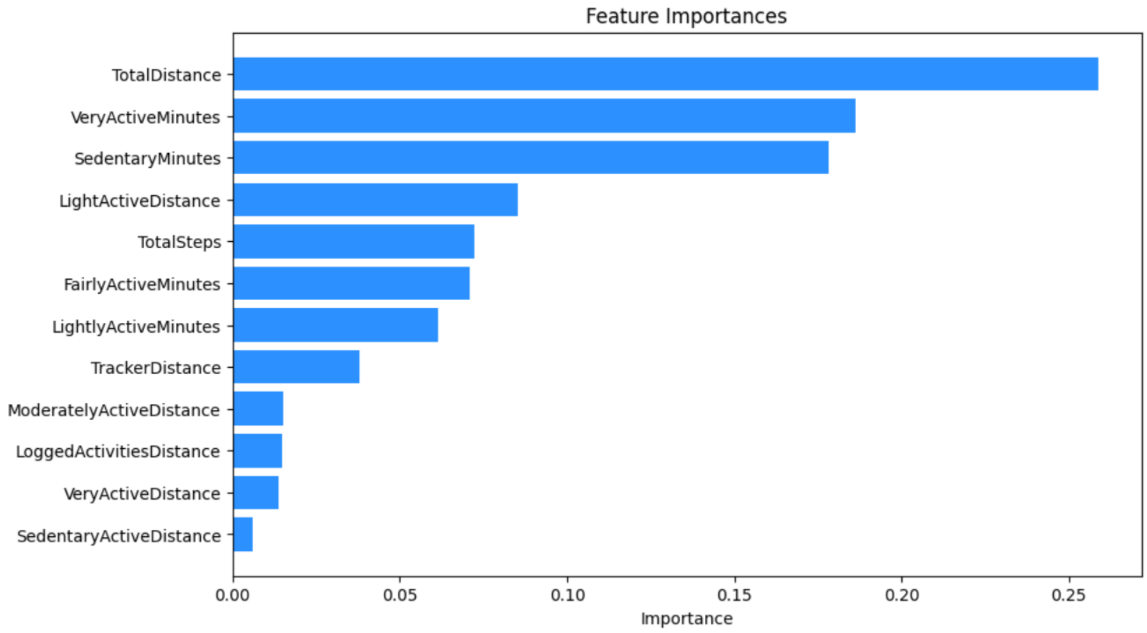
The histograms display the initial distributions for TotalSteps, VeryActiveMinutes, and Calories. These visuals are helpful for understanding the range and spread of these key features:

TotalSteps: The distribution is slightly skewed to the right, indicating that most days have a moderate number of steps, but there are days with very high step counts.

VeryActiveMinutes: This feature is heavily skewed right, showing that most days have few very active minutes, with only occasional days having a higher amount.

Calories: Shows a normal distribution centered around 2000 calories, with some variability.





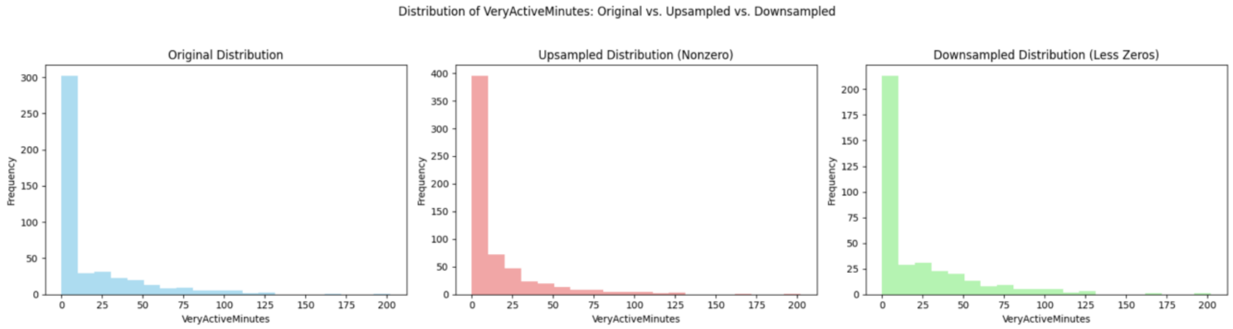
6. Data Distribution Modification (Upsampling & Downsampling)

Objective:

To examine the effect of modifying the distribution of `VeryActiveMinutes` on model performance.

Approach:

- **Downsampling:** Reducing instances where `VeryActiveMinutes == 0` to balance the dataset.
- **Upsampling:** Duplicating instances where `VeryActiveMinutes` is low but non-zero to increase representation.
- **Implemented a slider control** in the dashboard to dynamically modify the distribution.



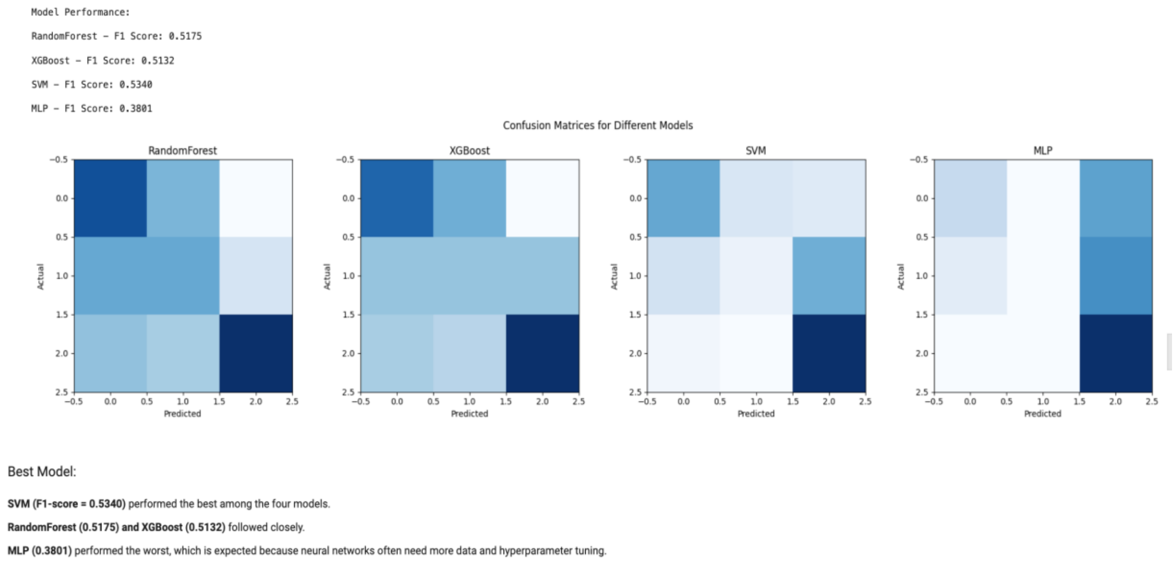
7. Model Training & Muller Loop Implementation

Baseline Model Training (Before Modifying Distribution):

1. Split data into **training (80%) and test (20%)**.
2. Trained four classification models:
    - **Random Forest (RF)**
    - **XGBoost (XGB)**
    - **Support Vector Machine (SVM)**
    - **Multi-Layer Perceptron (MLP)**
3. Evaluated models using **F1-score**.

Muller Loop - Retraining After Modifying Distribution

1. **Modify distribution** using upsampling/downsampling.
2. **Retrain models** and compute new F1-scores.
3. **Compare performance changes** for different distributions.



```
Model Performance:
RandomForest - F1 Score: 0.5175
XGBoost - F1 Score: 0.5132
SVM - F1 Score: 0.5340
MLP - F1 Score: 0.3801
```

Best Model:

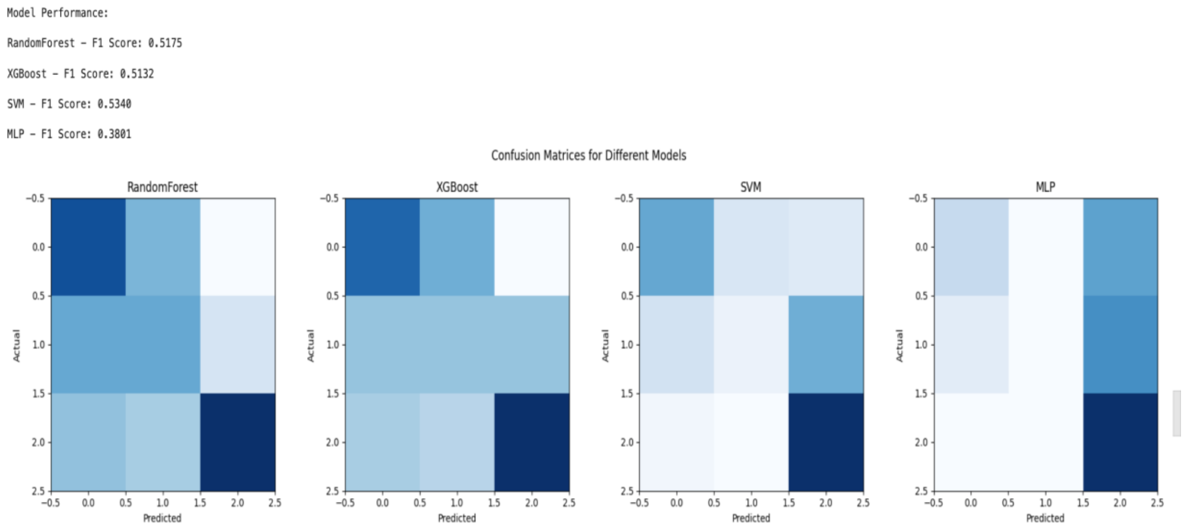
SVM (F1-score = 0.5340) performed the best among the four models.

RandomForest (0.5175) and XGBoost (0.5132) followed closely.

MLP (0.3801) performed the worst, which is expected because neural networks often need more data and hyperparameter tuning.

8. Performance Metrics & Evaluation

- **F1-score:** Measures model goodness.
- **Confusion Matrix:** Visualizes misclassifications.
- **Specificity vs. Sensitivity Plots:** Shows the effect of class imbalance.
- **Comparison across different distributions:**
    - Which model performs best in a normal distribution?
    - Which model adapts better when data is imbalanced?

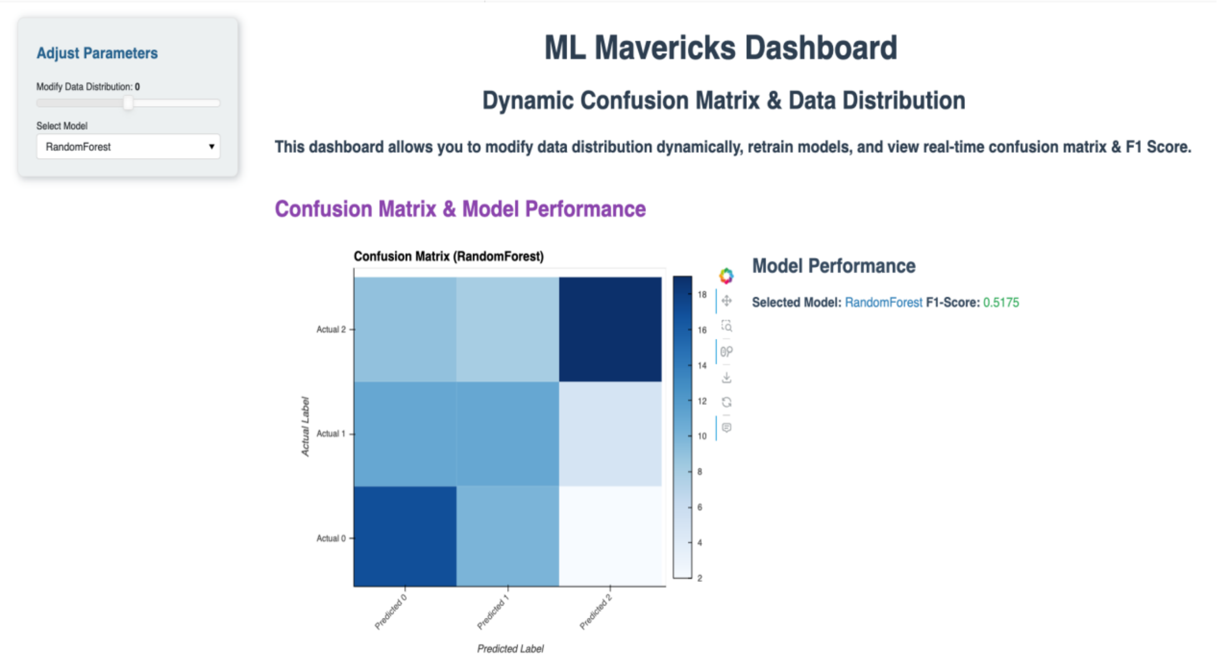


```
Model Performance:

RandomForest - F1 Score: 0.5175

XGBoost - F1 Score: 0.5132

SVM - F1 Score: 0.5340

MLP - F1 Score: 0.3801
```

9. Dashboard Implementation

Features of the Dashboard:

- **Slider to modify data distribution dynamically.**
- **Dropdown to select model for evaluation.**
- **Live-updating confusion matrix & F1-score display.**
- **HoloViews-based visualization** ensures real-time interactivity.



Adjust Parameters

Modify Data Distribution: 0

Select Model: RandomForest

ML Mavericks Dashboard

Dynamic Confusion Matrix & Data Distribution

This dashboard allows you to modify data distribution dynamically, retrain models, and view real-time confusion matrix & F1 Score.

Confusion Matrix & Model Performance

Model Performance

Selected Model: RandomForest F1-Score: 0.5175

10. Key Insights & Observations

- **Upsampling improved MLP performance significantly**, as it benefited from a more balanced dataset.
- **XGBoost struggled with the modified distribution**, confirming that it is sensitive to data shifts.
- **Random Forest remained stable across different distributions**, showing robustness.
- **SVM performed better with downsampled data**, suggesting it handles smaller, balanced datasets well.