Feature Engineering and Data Augmentation Report

Introduction

The dataset used for anomaly detection in crowds via smartphone data was enhanced through feature engineering and data augmentation techniques. This report documents the newly created features, their purposes, and the results from implementing statistical analysis to explore their relationships.

1. New Features Created

Several new features were engineered based on the existing accelerometer, gyroscope, and location data. These features help to capture more detailed information about movement, behavior, and anomalies in the dataset. Below is a detailed description of each feature:

1.1 Speed Change

- **Purpose**: Captures the change in speed between consecutive data points to detect acceleration and deceleration patterns.
- **Formula**:

```
python
Copy code
df['Speed_Change'] = df['Speed'].diff()
```

1.2 Direction Change

- **Purpose**: Identifies abrupt changes in the movement direction by calculating the difference in heading between consecutive points.
- **Formula**:

```
python
Copy code
df['Direction_Change'] = df['Heading'].diff().fillna(0)
```

1.3 Time Change

- **Purpose**: Computes the time difference between consecutive data points for time-sensitive feature engineering (e.g., rate of change in speed or acceleration).
- **Formula**:

```
python
Copy code
df['Time'] = pd.to_datetime(df['Time'], format='%H-%M-%S')
df['Time_Change'] = df['Time'].diff().dt.total_seconds().fillna(0)
```

1.4 Acceleration Rate

- **Purpose**: Measures the rate at which speed changes, helping to identify acceleration and braking events.
- **Formula**:

```
python
Copy code
df['Acceleration_Rate'] = df['Speed_Change'] / df['Time_Change']
```

1.5 Braking Intensity

- **Purpose**: Determines the intensity of braking by focusing on negative values of acceleration rate (i.e., deceleration).
- **Formula**:

```
python
Copy code
df['Braking_Intensity'] = df['Acceleration_Rate'].apply(lambda x: x
if x < 0 else 0)
```

1.6 Jerk

- **Purpose**: Calculates the rate of change of acceleration (jerk), which can be indicative of sudden movements or stops.
- **Formula**:

```
python
Copy code
df['Jerk'] = df['Acc_Magnitude'].diff() / df['Time_Change']
```

1.7 Cumulative Distance

- **Purpose**: Keeps a running total of the distance covered over time, aiding in trajectory analysis.
- **Formula**:

```
python
Copy code
df['Cumulative_Distance'] = df['Distance'].cumsum()
```

1.8 Speed Variance

- **Purpose**: Measures the variance in speed over a rolling window, providing insights into the steadiness of the movement.
- **Formula**:

```
python
Copy code
df['Speed_Variance'] = df['Speed'].rolling(window=5).var()
```

2. Time-Based Features

2.1 Rolling Mean of Accelerometer X (Acc X)

- **Purpose**: Computes the rolling average of the accelerometer's X-axis data over a 5-sample window, which helps smoothen out fluctuations and reveal underlying trends.
- **Formula**:

```
python
Copy code
df['Rolling_Mean_AccX'] = df['Acc X'].rolling(window=5).mean()
```

2.2 Moving Variance of Gyroscope X (Gyro X)

- **Purpose**: Calculates the variance of the gyroscope's X-axis over a 5-sample window to detect variations in rotational motion.
- **Formula**:

```
python
Copy code
df['Variance_GyroX'] = df['gyro_x'].rolling(window=5).var()
```

3. Engineered Feature

3.1 Total Acceleration

- **Purpose**: Computes the total magnitude of acceleration from the X, Y, and Z components of the accelerometer data, providing a comprehensive measure of motion intensity.
- **Formula**:

```
python
Copy code
df['Total_Acc'] = np.sqrt(df['Acc X']**2 + df['Acc Y']**2 + df['Acc
Z']**2)
```

4. Feature Importance Analysis

A correlation matrix was computed to identify the relationships between the various features. Significant observations include:

- **Speed** is positively correlated with **Heading** and **Cumulative Distance**.
- **Total Acceleration** is strongly correlated with **Acceleration Magnitude**, suggesting that these features provide similar information regarding motion intensity.
- **Jerk** is moderately correlated with **Acceleration Magnitude**, which is expected as it measures the rate of change in acceleration.
- Features like **Braking Intensity** and **Speed Variance** exhibit low correlations with most other features, indicating their unique contribution to the dataset.

5. Results Summary

The new features provide deeper insights into the behavior of the data and allow for more advanced analysis of movement, including sudden stops, changes in direction, and speed variations. The combination of time-based rolling statistics and engineered features will

improve anomaly detection, leading to better performance in security and crowd management applications.

6. Conclusion

The feature engineering and data augmentation efforts have enhanced the dataset by creating more informative variables that better represent the underlying dynamics of movement. These features will be useful for training machine learning models for real-time anomaly detection and understanding crowd behavior in various environments.

7. Saved Dataset

The updated dataset with all new features has been saved as `augmented_dataset.csv`.