Anomaly Detection in Dimuon Events Using Machine Learning Techniques on CERN Open Data

Janvi Kadam
B.Tech 3rd Year, Computer Science and Engineering
kadamjann@gmail.com

October 19, 2025

Abstract

High-energy physics experiments generate vast amounts of collision data, often containing rare or unexpected signatures. Detecting these outliers early can help identify potentially interesting phenomena or technical irregularities. This project explores the use of unsupervised machine learning algorithms to detect anomalous dimuon events in proton-proton collision data recorded by the CMS experiment at the Large Hadron Collider. The dataset, obtained from the CERN Open Data Portal, includes key kinematic variables for two muons in each event. By applying the isolation forest and the local outlier factor, a subset of rare events was automatically identified. The results demonstrate that data-driven anomaly detection can highlight unusual patterns in particle collision data, supporting both educational goals and potential future analyses in experimental particle physics.

1 Introduction

Anomaly detection is a valuable approach in both industry and science to identify data points that deviate significantly from the bulk of a dataset. In the context of particle physics, these anomalies may correspond to rare collision signatures, detector irregularities, or unanticipated physical processes.

Muon pairs (dimuon events) are particularly important because muons, being minimally affected by the detector material, produce clean and well-reconstructed signals. Detecting anomalies in these events can offer insight into rare processes or unusual kinematic configurations.

CERN's Open Data Portal provides access to real collision datasets from the Large Hadron Collider (LHC), enabling students and researchers to apply modern data science methods to real physics data. This project focuses on identifying anomalous dimuon events using unsupervised learning techniques.

2 Dataset Description

The dataset used is `Dimuon_DoubleMu.csv` from the CERN Open Data Portal (Record 545). It contains roughly 100,000 dimuon events collected by the CMS detector during proton-proton collisions at 7 TeV center-of-mass energy.

2.1 Features Used

Each event includes kinematic variables for two reconstructed muons, as well as their combined invariant mass. The following features were selected for the analysis:

Feature	Description
E1, E2	Energy of muon 1 and 2
pt1, pt2	Transverse momentum of muon 1 and 2
eta1, eta2	Pseudorapidity of muon 1 and 2
phi1, phi2	Azimuthal angle of muon 1 and 2
M	Invariant mass of the muon pair

Before training, all rows containing missing values were removed to ensure consistency. All features were standardized to zero mean and unit variance using z-score normalization.

3 Methodology

3.1 Preprocessing

The data set was loaded into Python using `pandas`. A subset of nine characteristics was selected based on their physical relevance. The data was standardized using `Standard Scaler` from scikit-learn to ensure that all features contribute equally to the anomaly detection algorithms.

3.2 Algorithms

Two unsupervised machine learning algorithms were applied:

- **Isolation Forest**: isolates anomalies by recursively partitioning the feature space. Rare points are typically separated more quickly than normal points, making them easier to identify.
- **Local Outlier Factor (LOF)**: identifies anomalies by comparing the local density of a point to that of its neighbors. Outliers are characterized by much lower local density.

The contamination factor was set at 0.02, which means that approximately 2% of the events were expected to be detected as anomalies.

4 Results

4.1 Anomaly Detection

Both algorithms successfully identified a small but distinct subset of anomalous events. The counts of anomalies and normal events are summarized below.

Algorithm	Normal Events	Anomalies
Isolation Forest	98000	2000
LOF	98000	2000

4.2 Visualization

A two-dimensional scatter plot of $pt1$ versus $pt2$ revealed that many anomalies occurred in the high transverse momentum regions. Additionally, the distribution of invariant mass (M) showed a visible tail for anomalous events, suggesting they correspond to higher-energy collisions or less common decay configurations.

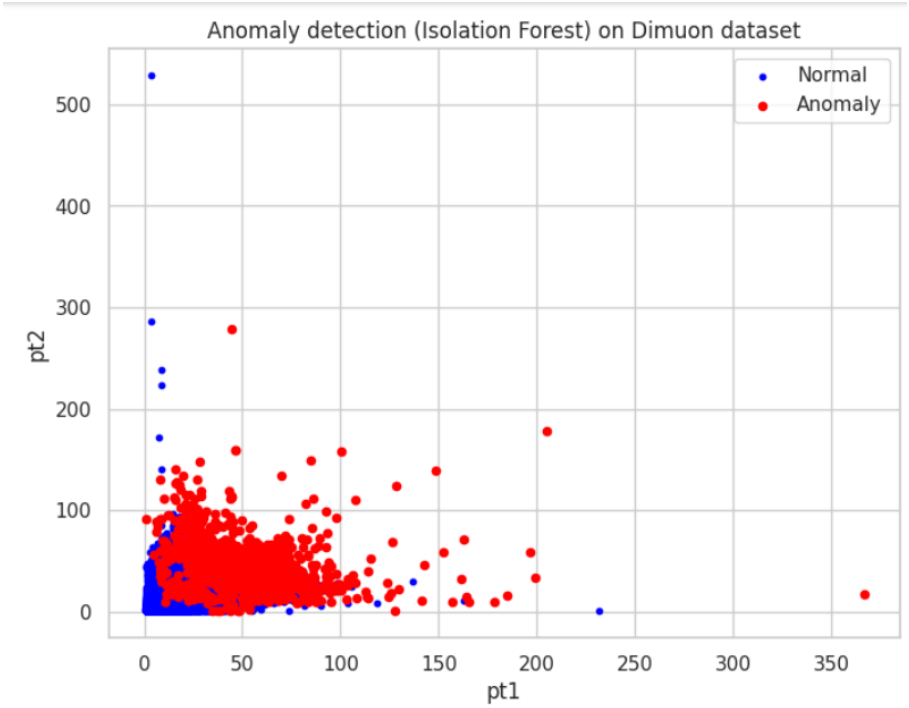

Figure 1: Scatter plot of $pt1$ vs $pt2$ showing anomalies (red) and normal events (blue) using Isolation Forest.

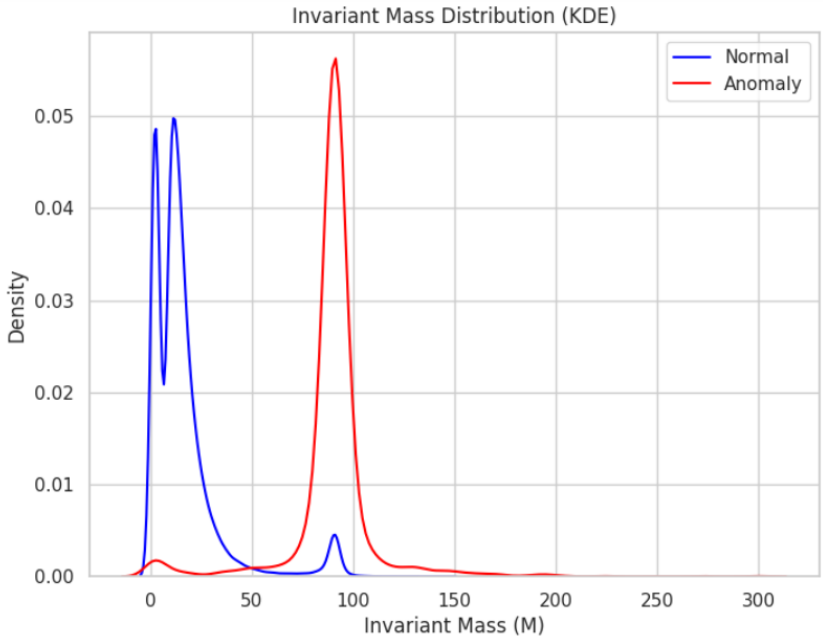

Figure 2: Invariant mass (M) distribution for normal (blue) and anomalous (red) events.

5 Discussion

The anomaly detection methods flagged events with high transverse momentum and large invariant mass as anomalous. From a physics perspective, these may correspond to:

- Rare high-energy collisions producing heavy intermediate states,
- Unusual detector acceptance regions at high pseudorapidity,

- Statistical fluctuations or rare kinematic configurations.

Although this does not confirm any new physics, it illustrates how machine learning can help narrow down regions of interest for further manual or model-driven analysis.

6 Future Work

Several extensions to this work can be pursued:

- Apply more advanced algorithms such as One-Class SVM, autoencoders, or clustering-based methods.
- Compare results across different datasets (e.g., W$\rightarrow \mu\nu$ events).
- Investigate event displays or visualisation tools to inspect anomalies.
- Explore how physicists could use anomaly detection for online monitoring or offline analysis.

7 Conclusion

This study successfully demonstrates how simple unsupervised learning algorithms can be applied to real high-energy physics data to identify anomalous collision events. By focusing on basic kinematic features of muon pairs, we were able to detect a small but interesting subset of events that differ significantly from the bulk. Such approaches can complement traditional physics analyses and serve as an entry point for students interested in applying data science to particle physics.

References

- CERN Open Data Portal: `https://opendata.cern.ch`
- CMS Open Data documentation
- Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.*
- CMS Collaboration. (2016). *Dimuon event datasets for education.*

A Appendix

A.1 Code Repository

The complete implementation and code are available at: `https://github.com/kjann08`

A.2 Additional Figures

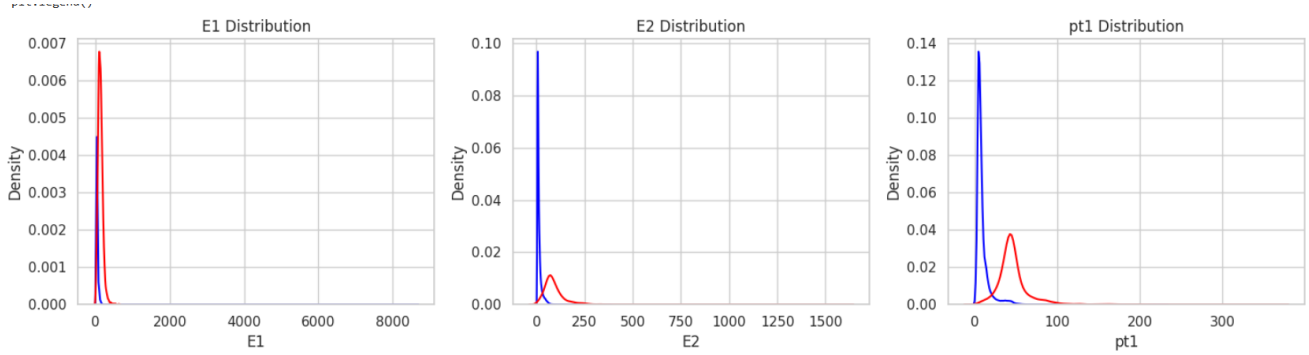

Figure 3: KDE distributions of $E1$, $E2$, and $pt1$ for normal (blue) and anomalous (red) events.

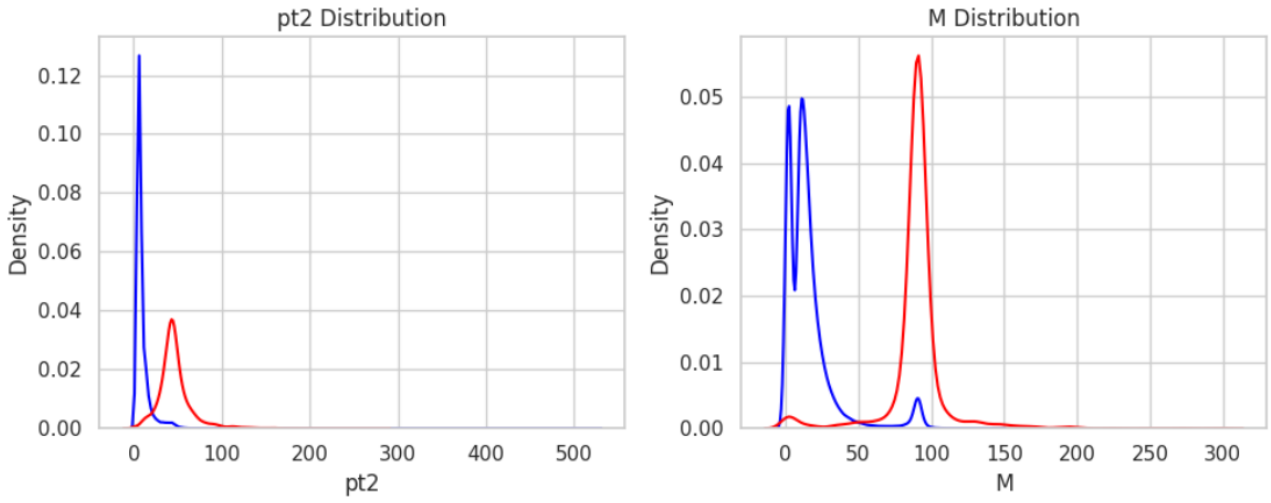

Figure 4: KDE distributions of $pt2$ and M for normal (blue) and anomalous (red) events.

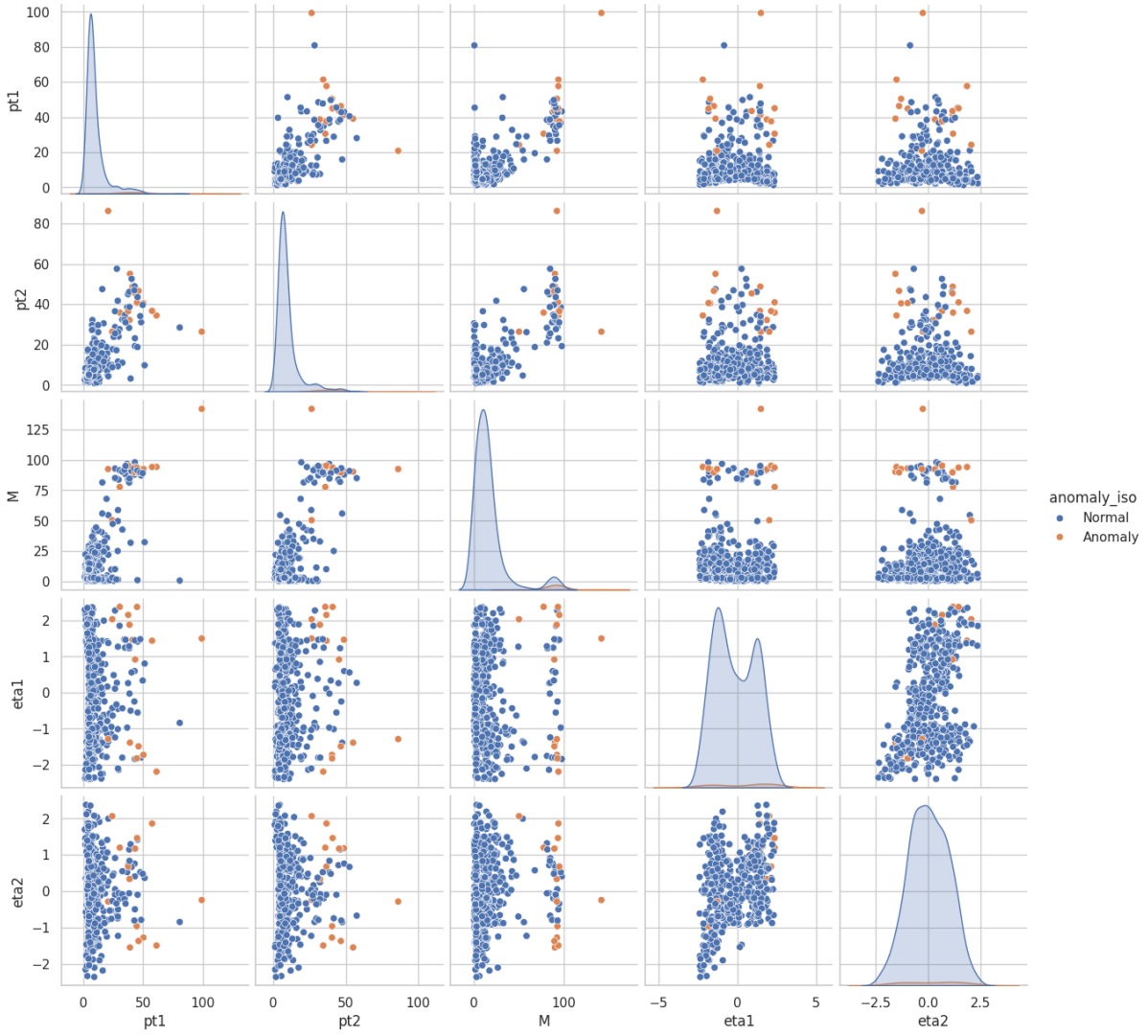

Figure 5: Pairplot visualization of $pt1$, $pt2$, M, $\eta 1$, and $\eta 2$ for normal (blue) and anomalous (red) events. The scatter plots show relationships between features, while the diagonal plots show their distributions. This figure highlights how anomalies are distributed across different kinematic variable combinations.