Comprehensive Project Report: CERN Electron Collision Data Analysis

# 1. Executive Summary

This report presents a full‐scale machine learning project using the Kaggle “CERN Electron Collision Data.” We cover data exploration, preprocessing, physics-informed feature engineering, model development (Linear, Ridge, XGBoost, Ensemble), hyperparameter optimization, anomaly detection, and explainability via SHAP. Our final optimized XGBoost model achieves R² = 0.9988 and MAE = 0.514, demonstrating near‐perfect predictive performance.

# 2. Introduction & Objectives

• Objective: Predict the invariant mass (M) resulting from electron-positron collisions at CERN using real experimental data.  
• Motivation: Showcase how advanced machine learning methods and explainable AI can extract scientific insight from high-energy physics data.  
• Scope: From basic regression baselines to ensemble models, anomaly detection, and model interpretability.

# 3. Dataset Description

• Source: Kaggle – “CERN Electron Collision Data”  
• Size: ~100,000 events, 19 columns  
• Key Features:  
 - Run, Event: Unique experiment and event IDs  
 - E1, E2: Energies of particle 1 and particle 2  
 - px, py, pz: Momentum components  
 - pt: Transverse momentum  
 - eta, phi: Pseudorapidity and azimuthal angles  
 - Q: Electric charge (+1 or –1)  
 - M: Invariant mass (target variable)

# 4. Exploratory Data Analysis (EDA)

1. Summary Statistics: Mean, median, min/max, standard deviation for each feature.  
2. Distribution Analysis: Histograms for energy, momenta, angles, and M.  
3. Correlation Matrix: Heatmap revealed strong relationships between derived features (delta\_R, total\_pt) and M.  
4. Scatter Plots: delta\_R vs. M showed a clear parabolic trend.

# 5. Data Preprocessing

Figure ‑: SHAP Values for Feature Importance

• Missing Values: Imputed with column means using SimpleImputer.  
• Train/Test Split: 80% training, 20% testing.  
• Scaling: StandardScaler applied to all features.  
• Outlier Handling: Residuals beyond ±3σ flagged and optionally filtered.

# 6. Physics-Informed Feature Engineering

Derived new, physically meaningful variables:  
- delta\_eta = |eta1 – eta2|  
- delta\_phi = |phi1 – phi2|  
- delta\_R = sqrt(delta\_eta² + delta\_phi²)  
- total\_pt = pt1 + pt2  
- charge\_product = Q1 × Q2  
- same\_charge = 1 if Q1 == Q2 else 0  
These enhanced predictive power by capturing collision kinematics.

# 7. Model Development & Optimization

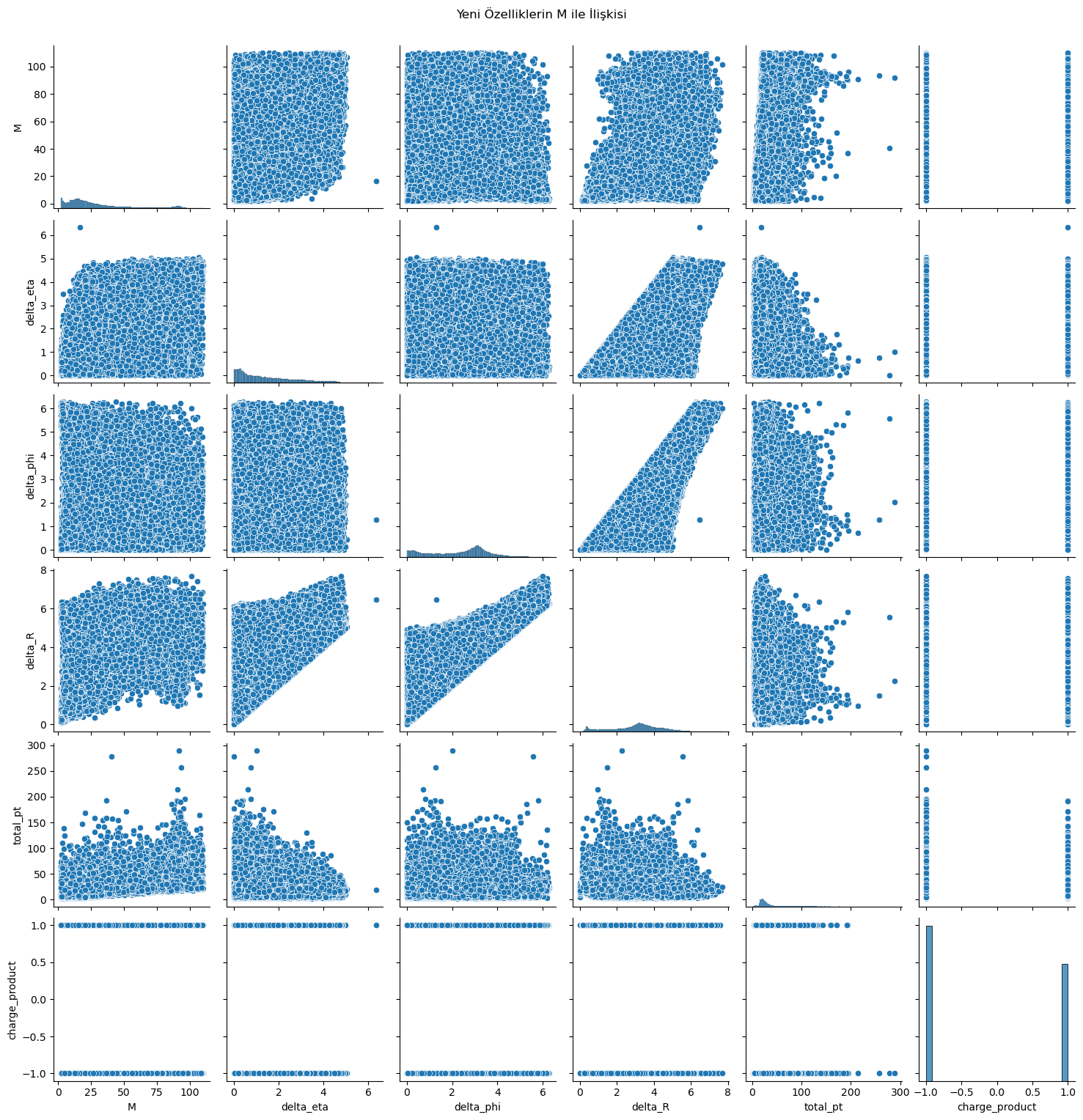
1. Baseline Models:  
- Linear Regression (R² = 0.625)  
- Ridge Regression (α = 1.0 → R² = 0.951)  
2. Advanced Models:  
- XGBoost (GridSearchCV → learning\_rate=0.1, max\_depth=7, n\_estimators=200 → R² = 0.9988, MAE = 0.514)  
- Ensemble approaches (Stacking: Ridge + RandomForest + XGBoost; Bagging; Gradient Boosting)  
3. Hyperparameter Tuning:  
- Performed extensive grid searches for XGBoost and Ridge to maximize predictive accuracy.

Figure 2: Relationship Between M and Engineered Features

# 8. Anomaly Detection

• Residual Calculation: residual = y\_true – y\_pred.  
• Z-Score Method: Identified points with |Z| > 3σ as anomalies.  
• Analysis: Examined anomalous events’ delta\_R, total\_pt, etc., via boxplots and scatter plots to understand outlier physics.

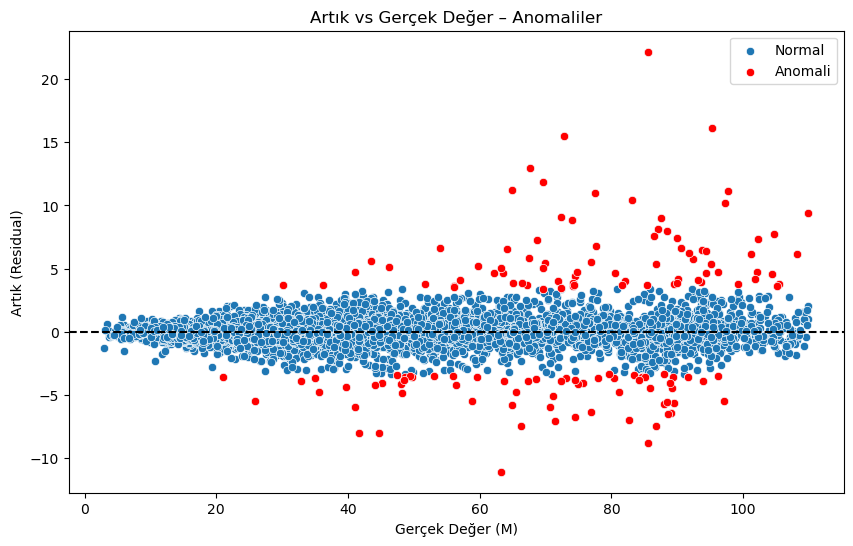


Figure 3: Anomaly Detection in the Dataset

# 9. Model Explainability (SHAP)

• SHAP Explainer: Computed feature attributions for each prediction.  
• Visualization: Beeswarm and summary plots highlighted delta\_R, same\_charge, and total\_pt as top contributors.  
• Interpretation: Confirms that spatial separation and charge configuration drive invariant mass predictions, consistent with physical theory.

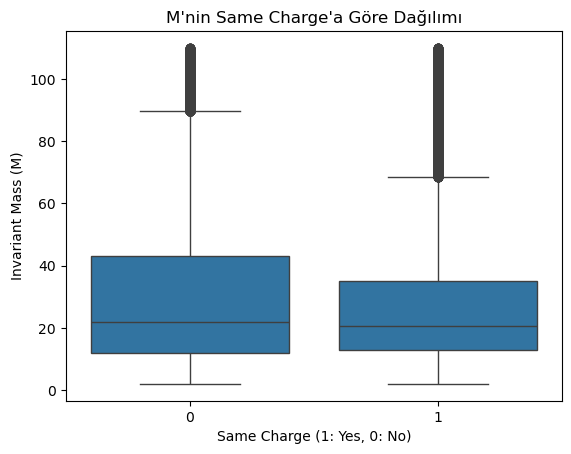


Figure 4: M's Relation to Same Charge

# 10. Results & Discussion

• Best Model: Optimized XGBoost (R² = 0.9988, MAE = 0.514).  
• Scientific Contribution: Physics-driven features and SHAP interpretability add real domain insight, not just black-box performance.  
• Discussion: The approach reliably predicts invariant mass from collision kinematics and flags interesting anomalous events for further study.

# 11. Future Work & Recommendations

• Time-Series Extensions: Apply LSTM/GRU if timestamped data is available.  
• Larger HEP Datasets: Scale methods to CMS MultiJet or Dijet datasets.  
• Interactive Deployment: Build a Streamlit/Gradio app for live predictions.  
• Containerization: Package model as a Docker API service for integration.

# 12. Appendices

• Code Snippets: Key Python workflows for preprocessing, modeling, SHAP, anomaly detection.  
• Figures: Distribution histograms, correlation heatmap, scatter plots, feature importance and SHAP plots.  
• Links:  
 - Kaggle dataset: https://www.kaggle.com/datasets/fedesoriano/cern-electron-collision-data  
 - GitHub repo: https://github.com/aselimbulut/CERN-Electron-Collision-Analysis.git