

# Electron Collision Mass Prediction

**P:30** Satyajeet Patil, Varad Patwardhan, Varad Sawant, Shubham Vasudeo Desai  
Department of Computer Science, North Carolina State University  
Raleigh, NC 27695  
{smpatil2,vspatwar,vsawant,sdesai8}@ncsu.edu

## 1 Data Set

The dataset used in the project was obtained from the CERN Open Data Portal. This dataset contains 100k dielectron events in the invariant mass range 2-110 GeV along with subset of the total event information.

## 2 Project Idea

Electron collisions are fundamental phenomena that are important for understanding atomic and molecular structure and other fields in physics. We will train machine learning models using various techniques. This project would involve data preprocessing, feature selection, and the development of a predictive model. Using this there can be improvements in anomaly detection, simulation of atom collision as well as experiment planning, and discovering new physics.

## 3 Software to Write

In this project we will be using Python for coding, along with that we will be using libraries like Scikit Learn and Google Colab for hosting the Jupyter Notebook and training the models.

## 4 Relevant Papers

- [1] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: <https://doi.org/10.1023/A:1010933404324>.
- [2] Adith Thaniserikaran et al. “The prediction of cern electron mass collision by using CAT-Boosting and LGBMR”. In: *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 2022, pp. 1–5. DOI: 10.1109/ICCCNT54827.2022.9984588.

## 5 Division of Work

Data preprocessing and Decision tree will be handled by Satyajeet Patil. Varad Patwardhan will handle LGBM Regressor and researching relevant papers for different models. Data visualization and Random forest will be done by Varad Sawant. Shubham Desai will perform relevant data selection and CATboost Regressor.

## 6 Midterm Milestone

We will complete the data preprocessing and visualization by the midterm. Also, we will train models to get results on models mentioned above which are already worked on before making more complex models.