NSSC 2025

NATIONAL STUDENTS' SPACE CHALLENGE

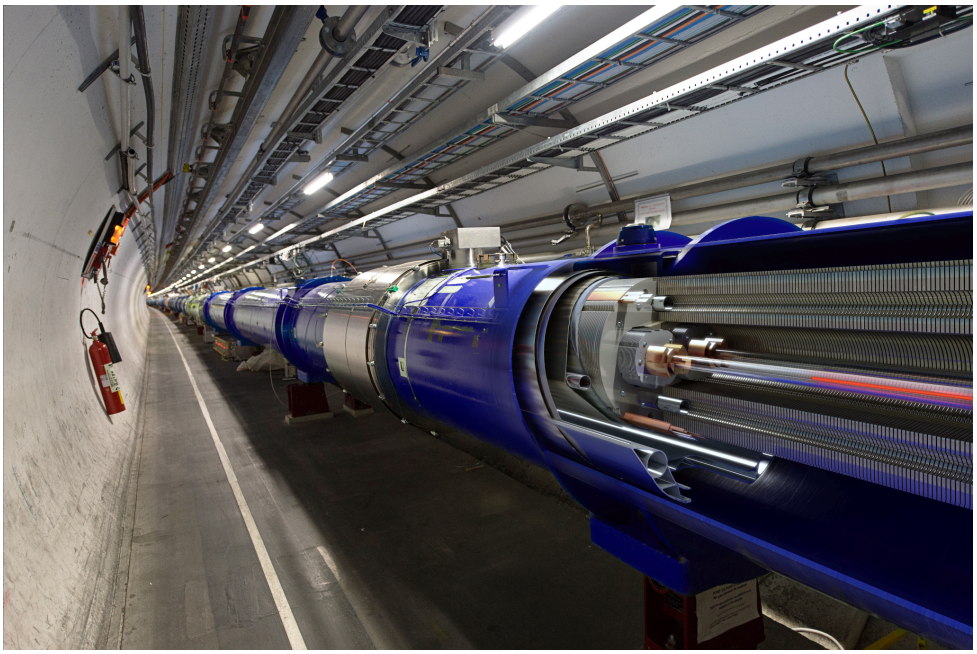
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

Data Analytics

General Instructions:

This is the Problem Statement of Data Analytics. The questions should be answered based on your thinking, and using any AI tools like ChatGPT, Gemini, Claude, etc., is strictly prohibited.

A plagiarism check will be done on all your answers, and violating any of these rules will lead to strict disqualification.



MOTIVATION:

Jets are streams of particles produced in high-energy collisions, such as those at the Large Hadron Collider (LHC). They provide vital insight into the fundamental forces of nature and can reveal rare processes or new particles beyond the Standard Model.

Studying jets is essential not only for understanding known physics but also for discovering anomalies that may signal new phenomena. Unusual jets could point to exotic processes, detector effects, or entirely new physics.

The universe remains a profound mystery, and particle physics is one of the key tools for probing its hidden structure. By applying deep learning and anomaly detection to jet data, we aim to uncover subtle patterns that traditional methods may miss—mirroring humanity's broader quest to decode the unknown and push closer to answers about the cosmos itself.

OBJECTIVE OF THE ANALYSIS:

By conducting this analysis, participants will not only learn to apply machine learning in a high-dimensional physics setting but also gain an understanding of how classification and anomaly detection techniques can contribute to real-world particle physics experiments. These methods may aid in the search for **rare events, exotic particles, or unknown cosmic phenomena**, reflecting the broader scientific effort to unravel the universe's mysteries.

RULES AND REGULATIONS:

- The problems are based on the **HLS4ML LHC Jet dataset**, which contains both image-like jet representations and tabular jet features.

- The Dataset contains two folders: one for training and another for testing. The training folder contains 61 files, and the validation folder contains 27 files. Each file contains 10,000 jet images, tabular data of 10,000 jets and a list of target classes for all 10,000 jets (target class is the same for both the corresponding image and tabular data point).

- The Feature Description provides a detailed description of the features.

- The problem statement consists of 4 major parts, each with multiple subproblems. The **weightage** of each question is mentioned next to it.

- Participants may use **any programming language or library** (e.g., Python with TensorFlow, PyTorch, scikit-learn). However, it is preferable to use **Google Colab or Jupyter Notebook** as the coding environment.

- Compile all the files in a structured manner and upload them on a drive folder. The drive link should be submitted through unstop. **Ensure that the viewer's access is granted.**

- Submissions should include:

    - A well-documented .ipynb notebook with code and outputs.

    - A **PDF report** containing detailed explanations, plots, tables, and results. All figures should be clearly labelled, and final answers should be highlighted.

- Each .ipynb file must have **markdown headings** to clearly indicate which code corresponds to which question and sub-question. Failure to organise code properly may lead to disqualification.

- Teams of **2 to 4 participants** are allowed. Collaboration and division of tasks are encouraged.

- Each team will be given **15 minutes to present** its solution. Presentations should be clear, concise, and emphasise insights gained from the analysis.

- Evaluation criteria will include:
    - **Correctness and completeness** of the solution.
    - **Clarity of explanation** and logical flow.
    - **Innovation and creativity** in feature engineering, model design, or analysis.
    - **Interpretation of results** in the context of physics.

The problem statement carries a total of **100 MARKS**. Your score will be scaled to **60 MARKS**, with the remaining **40 MARKS** allocated to the presentation.

PROBLEM STATEMENT

1. Data Preprocessing (20 points)

For this part, you can take any one file. So you will have a dataset of 10000 records.

1.1 Feature Handling (5 points)

- Clearly separate **image features** (energy distributions) from **tabular features** (physics-derived jet variables).
- Identify missing values in the tabular dataset.
- Apply imputation strategies for numerical features (e.g., mean, median, interpolation).
- Justify your choice and discuss how handling missing values may affect downstream models.

1.2 Tabular Data Preparation (5 points)

- Perform Principal Component Analysis (PCA) on the tabular dataset.
    - Compute the eigenvalues and eigenvectors of the covariance matrix.
    - Report the explained variance ratio for the top principal components.
    - Plot a **scree plot** to visualise variance captured per component.
- Comment on how PCA reduces dimensionality and noise while preserving important information.

1.3 Dataset Summary (5 points)

- Summarise the dataset: number of rows, number of features, image size, number of classes (if applicable).

- Provide **distribution plots** for selected numerical features (e.g., transverse momentum, energy).

2. Model Development (45 points)

You must use at least 251107 records for this part.

2.1 CNN on Image Data (25 points)

- Design and train a CNN classifier for jet images (you can also use generators, up to your choice, but specify your reasoning).
    - Clearly specify the architecture, including the number of convolutional layers, kernel sizes, activation functions, and pooling strategies. Alternatively, you can use a pre-trained model and then fine-tune it using your available data, whichever approach yields better accuracy.
    - Must use validation data and explain the amount of data chosen.
    - Justify design choices (e.g., why ReLU or why a pre-trained model).

- Implement training checkpoints to save model weights and log performance metrics.

- At the end of each epoch, report the training and validation accuracy and loss explicitly.

- Plot accuracy and loss curves across epochs. Analyse whether the model is converging, overfitting, or underfitting.

- Visualise CNN feature maps and activation layers for selected jets.

- Display and analyse 10 misclassified jets (if there are fewer than 10, use all of them), comparing predicted vs. true labels.

Hint: You can go through a research paper and use its architecture and methods if needed, but clearly specify anything like that if used.

2.2 Models on Tabular Data (20 points)

- Train a baseline model of your choice (e.g., Logistic Regression, Random Forest, or MLP) on the **raw tabular dataset**, selecting the one that you think is most suitable for it. Explain the reasoning behind your choice.
    - Report performance metrics: accuracy, precision, recall, F1-score, and ROC AUC.
- Train the same model on the **PCA-reduced dataset**.
    - Report the same metrics and compare against the baseline.
- Discuss the trade-offs: Does PCA improve accuracy, reduce overfitting, or make training faster?
- Provide **comparison tables and graphs** for the before-and-after PCA results.

3. Model Comparison (15 points)

- Compare the CNN image-based model with the best-performing tabular model (before vs. after PCA).
- Present metrics side by side: accuracy, precision, recall, ROC AUC, and confusion matrices.
- Plot ROC curves for CNN and tabular models on the same graph for direct comparison.
- Discuss which modality (image or tabular) provides stronger predictive power and why.
- Relate findings to the physics perspective: images capture spatial distributions, while tabular features capture derived physical parameters.

4. Anomaly Detection (20 points)

4.1 CNN Autoencoder for Anomaly Detection (15 points)

- Build a CNN autoencoder trained on the jet image dataset.
    - Specify architecture: encoder and decoder layers.
    - Justify the bottleneck size (latent dimension).
- Train the autoencoder to reconstruct “normal” jets.
- For each test image, calculate reconstruction error (e.g., Mean Squared Error).
- Use reconstruction error as an anomaly score.
    - Plot the distribution of errors.
    - Propose a threshold for anomaly detection (e.g., mean + 2σ).
    - Justify the choice and discuss the trade-off between sensitivity and specificity.

4.2 Visualisation and Analysis (5 points)

- Visualise the top 5 jets with the highest reconstruction errors alongside their reconstructions.
- Print the total number of anomalies detected.
- Discuss how anomalies might correspond to rare physics processes in real-world experiments (e.g., new particles, detector noise, or beyond-Standard-Model signals).