|  |  |  |  |
| --- | --- | --- | --- |
| **Thesis Progress Form** | | | |
|  | | | |
| **M.TECH. IN INFORMATION TECHNOLOGY**  **CENTRE FOR IT EDUCATION**  **P.G.DEPARTMENT OF STATISTICS, UTKAL UNIVERSITY.** | | | |
| **Candidate’s Class Roll No:** | 21M.Tech.IT 005 | **Candidate’s EXAM Roll No:** | 412VMIT 21006 |
| **Candidate's Name:** | Bibhu Prasad Mohanty | | |
| **Thesis Title:** | Charm-Jet Physics with Machine Learning | | |
| **Supervisor(s) details with Institute Address & Contact Number:** | Mrs. Sonia Parmar  Ph.D , Punjab University  Contact Number-99515811588 | | |
| **Secondary Supervisor:(if any)** |  | | |

***Signature of the Student Signature of the Supervisor(s)***

**ABSTRACT :**

The study of charm jets is of great importance in high-energy physics, as they offer valuable insights into the behaviour of heavy quarks in proton-proton collisions. In this thesis, we focus on generating proton-proton collision events using various hadronic models, including PYTHIA. We reconstruct charm jets by clustering charged particles with the anti-kt reconstruction algorithm and identifying a D0 meson as a constituent in each jet. We then use machine learning techniques to analyse the properties of these charm jets and explore their behaviour in different collision scenarios. Our results demonstrate the power of machine learning in studying complex physical phenomena and provide new insights into the behaviour of heavy quarks in high-energy collisions. This research has significant implications for the ongoing quest to understand the fundamental nature of matter and the universe.

Keywords - charm, jets, machine learning, D0 meson, heavy-flavour jets, heavy quarks

**INTRODUCTION:**

The study of particle physics aims to understand the fundamental nature of matter and the universe. High-energy collisions of protons provide an excellent laboratory for investigating the behaviour of tiny subatomic particles like quarks and gluons. Quarks and gluons are indivisible fundamental particles that cannot be broken down into smaller components. The strong interaction, one of the four fundamental interactions like gravity and electromagnetism, is what binds quarks together to form protons, neutrons, and other hadron particles.

Jets, which are sprays of particles that result from the breakup of quarks and gluons, are particularly useful for understanding the strong interaction. Scientists have studied jets extensively in particle physics experiments, both at colliders like the Large Hadron Collider (LHC) and in other experiments.

Heavy quarks, such as charm quarks, are of particular interest due to their large masses, which can significantly affect the properties of jets that they produce. In particular, charm jets offer a unique opportunity to study the properties of heavy quarks in the context of the strong interaction. Charm jets are jets reconstructed using charged particles and containing a D0 meson as a constituent in each jet. The study of charm jets has been a topic of interest in both experimental and theoretical particle physics, as they provide insights into the production and decay of heavy quarks.

Machine learning techniques have become increasingly important in particle physics research. These techniques allow for the analysis of large datasets and complex physical phenomena, and have been used to identify particles, classify events, and improve the accuracy of theoretical predictions.

Machine learning involves developing algorithms and different statistical models that enable computers to automatically learn patterns from the data. Based on the newly learned knowledge from existing data, they can become more accurate at predicting new outcomes. In many aspects, machine learning is present in our day-to-day life, for example: recommendations we receive on YouTube channels while watching any program, as well as on online shopping platforms, and voice assistants on our smartphones. Machine learning is useful in almost every field, such as finance for fraud detection, health care for disease diagnosis, transportation for traffic prediction etc.

In this thesis, we use machine learning techniques to analyse the properties of charm jets and explore their behaviour in different collision scenarios.

1.2 Objectives

The main objective of this thesis is to study charm-jet physics using machine learning techniques. Specifically, we aim to:

* Generate proton-proton collision events using various hadronic models, including PYTHIA.
* Reconstruct charm jets using the anti-kt reconstruction algorithm and identifying D0 mesons as constituents in each jet.
* Use machine learning techniques to analyse the properties of charm jets and explore their behaviour in different collision scenarios.

**LITERATURE SURVEY:**

**METHODOOGY:**

2.1 PYTHIA

PYTHIA is a powerful tool widely used in high-energy physics research for simulating particle collisions and generating Monte Carlo event samples. It is specially designed for simulating proton-proton collisions at high energies. The input to PYTHIA includes the parameters of the proton-proton collision, such as the centre-of-mass energy, different stages of the colliding protons. PYTHIA then generates a series of events that simulate the different possible outcomes of the collision. By using different aspects of proton-proton collisions such as, properties of different particles, correlation between different particles, the energy and momentum of the produced particles, one can do a nice research.

2.2 Machine Learning

A. Linear Regression

Linear regression is a statistical method used to analyse the relationship between a dependent variable and one or more independent variables. It is a technique used to model and predict the behaviour of a continuous dependent variable based on one or more independent variables.

In a simple linear regression, there is only one independent variable and one dependent variable, and the relationship between them is modelled using a straight line. The goal of the analysis is to determine the equation of the line that best describes the relationship between the variables, which can be used to make predictions about the dependent variable based on the values of the independent variable.

In multiple linear regression, there are two or more independent variables, and the relationship between them and the dependent variable is modelled using a plane or higher-dimensional surface. This technique can be used to identify the relative importance of each independent variable in explaining the variation in the dependent variable.

Linear regression is widely used in various fields such as finance, economics, social sciences, engineering, and many others, where the goal is to analyse and predict the behaviour of a dependent variable based on one or more independent variables.

### B. Logistic Regression

Logistic regression is a statistical method used to analyse the relationship between a dependent variable and one or more independent variables, where the dependent variable is categorical or binary in nature. In other words, it is a technique used to model and predict the probability of an event occurring, based on the values of one or more independent variables.

In logistic regression, the dependent variable is represented by a binary variable that takes on one of two values, typically 0 or 1, and the relationship between the dependent variable and the independent variables is modelled using a logistic function. The logistic function transforms the values of the independent variables into a probability score, which can be interpreted as the likelihood of the dependent variable taking on the value 1.

The goal of logistic regression is to determine the coefficients of the independent variables that best predict the value of the dependent variable, by estimating the probability of the dependent variable taking on the value 1, given the values of the independent variables. This can be used to make predictions about the probability of an event occurring, based on the values of the independent variables.

Logistic regression is widely used in various fields such as medicine, biology, marketing, and many others, where the goal is to analyse and predict the probability of an event occurring, based on one or more independent variables.

### C. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the family of ensemble methods. It is a type of supervised learning algorithm that is used for both classification and regression tasks. Random Forest works by constructing multiple decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The individual decision trees are generated using a random selection of features and a random subset of the training data, hence the name “Random Forest”.

D. Decision tree

The decision tree is a machine learning algorithm used for classification and regression problems. It is a tree-structured model that recursively splits the training data into subsets based on the value of a selected attribute or feature until a stopping criterion is met. The decision tree algorithm creates the tree in a top-down manner and chooses the feature that best separates the data into different classes or values. Once constructed, the decision tree can make predictions on new data by traversing the tree based on the values of the new data's features. Decision trees have advantages such as handling both categorical and continuous data, simplicity and interpretability, and handling missing data. However, they can be prone to overfitting, and the choice of split criterion and stopping criterion can impact the accuracy of the model.

Broadly, there are two types into which the various machine learning techniques described above can be classified into.

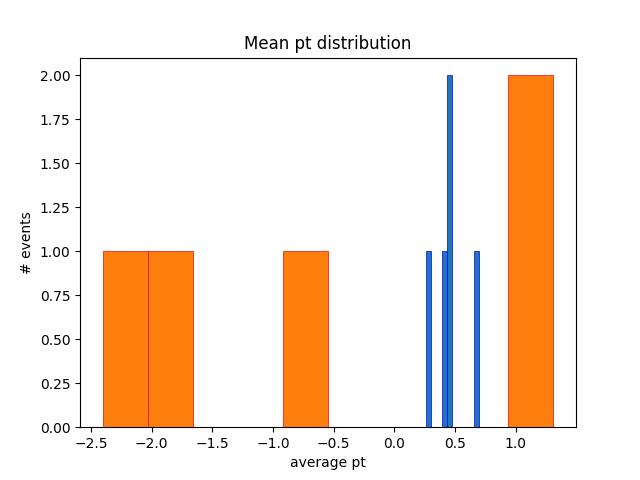
A. Unsupervised Learning

The biggest difference between unsupervised learning and supervised learning is absence of data labels in training. Training samples for unsupervised learning have no labels and no definite results for output, the computer needs to learn the similarity between samples by itself and classify the samples. The advantage of unsupervised learning is that there is no need to label, reducing the influence of human subjective factors on the results.

B. Supervised Learning

Supervised learning needs the labels of the training data. Common supervised learning algorithms include logistic regression, Naive Bayesian, Support Vector Machine, artificial neural network and random forest. They use the self-learning characteristic of neural networks to transform cracks recognition into crack probability judgement of each sub-block image in the work.

**EXPERIMENTAL RESULTS AND DISCUSSION :**



**CONCLUSION:**

In conclusion, this master's thesis aims to use machine learning techniques to study charm jets and explore their behaviour in different collision scenarios. The study of charm jets is of great interest in both experimental and theoretical particle physics, as they provide insights into the production and decay of heavy quarks. The methodology includes the use of PYTHIA, a powerful tool used for simulating particle collisions and generating Monte Carlo event samples. Additionally, linear and logistic regression techniques will be used to analyse the properties of charm jets and identify the relative importance of independent variables in explaining the variation in the dependent variable. The study of charm-jet physics using machine learning techniques is expected to provide new insights into the properties of heavy quarks and the strong interaction. The results of this research may have implications for the design and interpretation of future particle physics experiments, and could lead to a better understanding of the fundamental nature of matter and the universe.