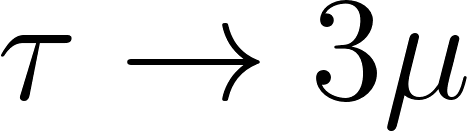
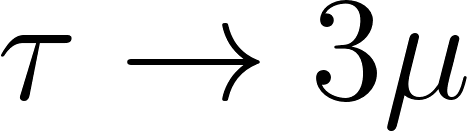
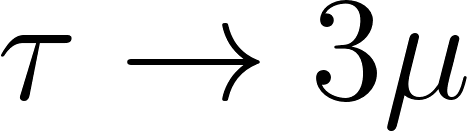
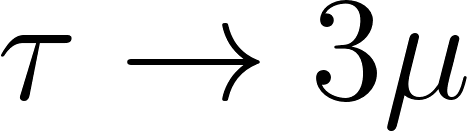
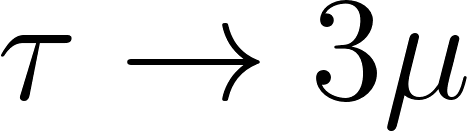
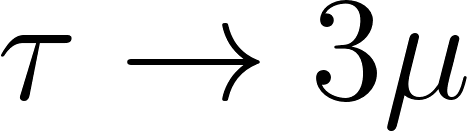
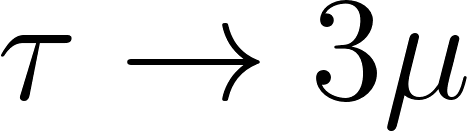
Flavours of Physics: case study

For source of the data and more info: <https://www.kaggle.com/c/flavours-of-physics>

# **Overview**

1. The case study named “Flavours of Physics” is a very interesting problem regarding search for “New Physics”. By “New Physics” it is meant that a new fundamental theory of physics is being searched since our existing understanding of physics or science is not enough in explaining some observed phenomenon in nature. One such observed phenomenon is asymmetry in the existence of matter-antimatter. It is well known that nature prefers symmetry in all phenomena. But in spite of rigorous search by mankind there has not been any evidence of enough antimatter in the universe as much as there is matter. All matter and antimatter are made up of some fundamental particles and these particles interact with each other via 4 fundamental forces of nature. These forces are basically communicated via some fundamental particles. The theory that explains these particles, currently, is called the “Standard Model” (SM). The fundamental particles that make matter are of type Fermions and are divided into two kinds called Leptons and Quarks. On the other hand, the particles that communicate the fundamental forces of nature are of type Bosons. The particles that make matter exist in pairs of particles and antiparticles i.e. for every particle there exists an antiparticle. These matter particles keep contributing in reactions which are generally called “decays”. In decays one particle changes identity into several other particles. Even though our current understanding of these particles and their interactions according to SM gives us satisfactory explanation of several phenomena (The SM had even predicted existence of Higgs boson, the so called “god particle” way before its experimental observation in 2012 at Large Hadron Collider (LHC) at Geneva, Switzerland), it still fails to explain some phenomena (like matter-antimatter asymmetry). To explain such phenomena a new theory would be required which is being called “Beyond Standard Model” (BSM). So the goal of this case study is to train a Machine Learning/ Deep Learning (ML/DL) model which, given some parameters or variables obtained from an experiment, could successfully predict if a particular decay is possible for the given parameters or not. The phenomena under observation here is [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) decay where [](https://www.codecogs.com/eqnedit.php?latex=%5Ctau#0) and [](https://www.codecogs.com/eqnedit.php?latex=%5Cmu#0) are two of 12 fundamental particles that make matter and antimatter. This decay violates The rule of “Lepton Flavour Conservation” which is a fundamental rule for occurrence of decays in SM. Lepton flavour is a quantum number which needs to be conserved during a decay in order for that decay to happen (Consider lepton flavour to be like a charge on a particle). Each particle has its own value of lepton flavour. According to SM this decay is prohibited. And it has not been observed to happen experimentally too. But According to BSM theory (Which is not a complete theory yet but this theory can be considered as evolved SM which could explain many phenomena not explained by SM) this kind of decay is possible. So if such kind of decay is found experimentally then this would be a strong step forward towards BSM. For this purpose, the dataset provided has data points made up of several variables and the data has been acquired from proton-proton collision in one of the experiments at LHC. There are two events one is signal event which suggests the [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) decay has happened and other is background event suggesting [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) decay has not happened. Since [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) has not been observed experimentally yet, the signal events have been generated by theoretical simulations on the basis of other decays which are topologically similar to [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) decay. There is a train dataset on which the model is to be trained. There are two tests to be conducted to check if the model is not biased towards either of the true and simulated data points and to check non-correlation of the results with the mass of the particles. There is a test dataset for which signal/background events are not given, this is the dataset on which predictions are to be made for submission. So given a data point the model should predicts whether the [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) has happened for the given set of variables or not. Such a model could be very helpful in further contribution towards actually finding [](https://www.codecogs.com/eqnedit.php?latex=%20%5Ctau%20%5Crightarrow%203%20%5Cmu%20#0) decay and contribution towards building a complete theory of BSM which could be helpful to solve many of the mysteries of this universe.

# 

# **Research-Papers and reference materials:**

1. Flavours of Physics: the machine learning challenge for the search of τ − → µ −µ −µ + decays at LHCb

Link:https://storage.googleapis.com/kaggle-competitions/kaggle/4488/media/lhcb\_description\_official.pdf

This is the actual paper authored by researchers who put forward this challenge. In this paper they have presented the details of the “Flavours of Physics” problem. A brief introduction on theoretical physics regarding the concepts in this problems is given along with description of the data and evaluation procedure for testing the model. The basic information about the experimental setup of the particle accelerator at LHC is given along with detectors used to collect the data for this problem. As discussed earlier there are two events, signal(1) and background(0). Signal being happening of decay and background being absence of the decay. Data for background events has been collected by actual detectors at LHC and data for signal events has been generated via theoretical simulation. This is the target variable that is to be predicted. The evaluation is done in two steps: firstly, the classifier is checked not to depend too strongly on the discrepancies between real data and simulation and the classifier is checked not to be too correlated with the τ mass. The second step is the evaluation of the classifier using the weighted area under the ROC curve. There are four datasets described below:

1. training.csv is a labelled data set (the signal being 1 for signal events, 0 for background events) to use for training the classifier. Background events come from real data mass side-bands3 and from 3 The mass side-bands are mass regions around the τ mass region in which the signal 6 simulation.
2. check agreement.csv is a labelled data set (the signal being 1 for simulated data, 0 for real data) with the same features as in the training.csv. This data set is used to check the agreement between simulated and real data .
3. check correlation.csv is a data set with the same features as the training.csv, to check yourself correlation of the classifier with the τ mass as described .
4. test.csv is a non-labelled (signal and background are mixed and indistinguishable for the participants) data set, containing simulated signal events and real background data, simulated events and real data for the control channel (latter form the main part of the test data set).

2. The Standard Model of Physics

Link: <https://arxiv.org/ftp/arxiv/papers/1412/1412.4094.pdf>

This paper gives a historical account of development of the Standard Model of Particle Physics over the last few decades. This paper accounts for how different theories of physics describing different seemingly unrelated phenomena were unified to propose one single theory called Standard Model. Quantum electrodynamics, describing the interactions of electrons with light, was incorporated into the electroweak theory, describing electromagnetic and weak nuclear interactions. There are 4 fundamental forces of nature which are gravitational force, electromagnetic force, strong nuclear force and weak force. The SM incorporates all these forces except gravitational force into one single theory. The incorporation of gravitational force into this theory would be Grand Unification Theory(GUT) which would be moving toward Beyond Standard Model(BSM). The account of how these seemingly unrelated forces are incorporated into one theory of SM is given in this paper.

3. New approaches for boosting to uniformity

Link: https://iopscience.iop.org/article/10.1088/1748-0221/10/03/T03002/pdf

In this paper several methods for use of multiclass classifiers based on boosting have been discussed that are extensively being used in the problems of particle physics. This paper explores several novel boosting methods that have been designed to train a model based on boosting which takes in a variable from a multivariate space and gives an output depending on the problem at hand. Several approaches, e.g. uBoost, kNNAda, uGbkNN, uGBFL etc., are discussed along with their results. According to the findings of this paper, out of all these approaches, the uGBFL algorithm outperformed the other algorithms.

# 

# **First Cut Approach**

1. First of all the EDA of the training dataset will be done. Through EDA each variable of the dataset will be tested for its discriminating ability for distinguishing between two classes. Further correlation of each variable with other variables will be tested. Dataset will be examined for any imbalance. Number of variables will be reduced based on correlation of variables with each other. Once EDA is done a baseline model will be trained on the train dataset. After observing the results variables will be checked for their importance and unimportant variables will be emitted. Based on literature survey from various sources, new features will be engineered and another baseline model will be trained. Again unimportant variables will be emitted and then the best fitting algorithm(ML based or DL based) will be determined with hyperparameter tuning. Once the model is trained after hyperparameter tuning, the model will be tested for its non-dependence on simulated data using check\_agreement.csv dataset. The model will also be tested for its non-dependence on the mass of [](https://www.codecogs.com/eqnedit.php?latex=%5Ctau#0) lepton. These two are necessary conditions the model has to pass. Once the model passes these conditions, the performance of the model will be checked with the weighted AUC score. The performance of the model will be compared with baseline model and if the performance is satisfactory then predictions will be made on test data.