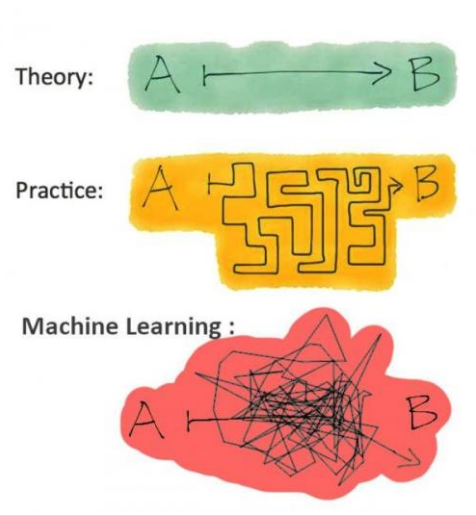
# Элементы машинного обучения в практикуме по вычислительной физике

Кирюхина Наталия Владимировна,
доцент кафедры физики и математики

Калужский государственный
университет им. К.Э. Циолковского



## Введение

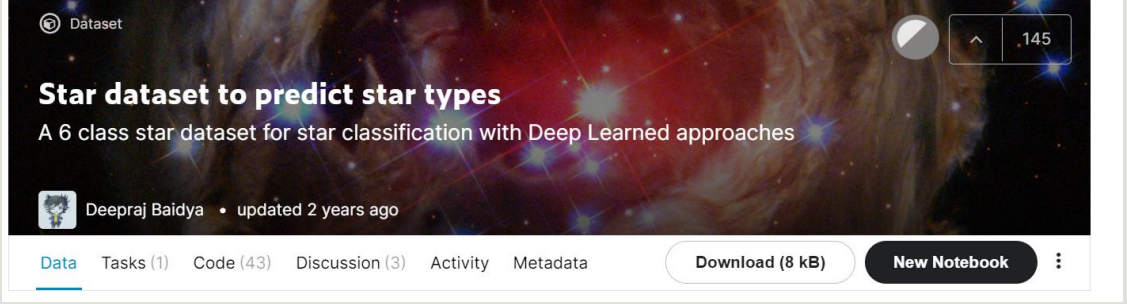
Алгоритмы интеллектуального анализа данных приобретают все большее значение в физическом эксперименте, особенно в области физики высоких энергий и астрофизики. Продемонстрировать возможности методов машинного обучения можно в рамках курса вычислительной физики.

**Цель:** разработка и практическая апробация системы учебно-исследовательских заданий для практикума по вычислительной физике на тему «Машинное обучение в современной физике и астрофизике».

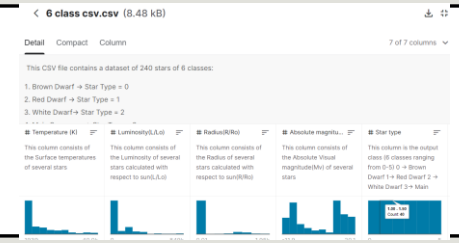
# Методология, методы и методики

| | |
|---|---|
| Выбор темы | • актуальность использования МО в области физики элементарных частиц и астрофизики |
| Отбор содержания | • мировоззренческое и методологическое значение тех проблем, на основе которых проектируются задания. |
| Объекты учебного исследования | • наборы данных, размещенные в открытом доступе на платформе Kaggle (площадка для соревнований по машинному обучению)<br>• требования к набору: небольшое число признаков с понятным для обучающихся предметным смыслом |
| Работа с данными | • разведочный анализ с использованием различных способов визуализации, разбор примеров постановки и решения задач для данного набора |

**Пример задачи для начинающих: разведочный анализ данных и классификация**



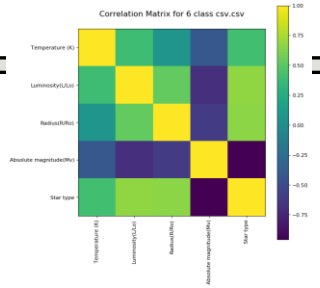

https://www.kaggle.com/deepu1109/star-dataset

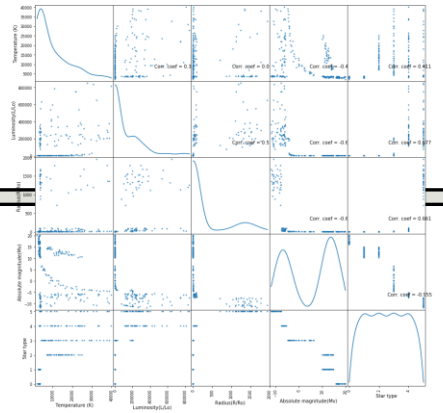



**Задание:**

**1. Построить диаграммы  распределения по температуре, относительной светимости, радиусу и другим признакам для всех типов и для каждого типа в отдельности.**

**2. Построить графики рассеяния для каждой пары признаков.**



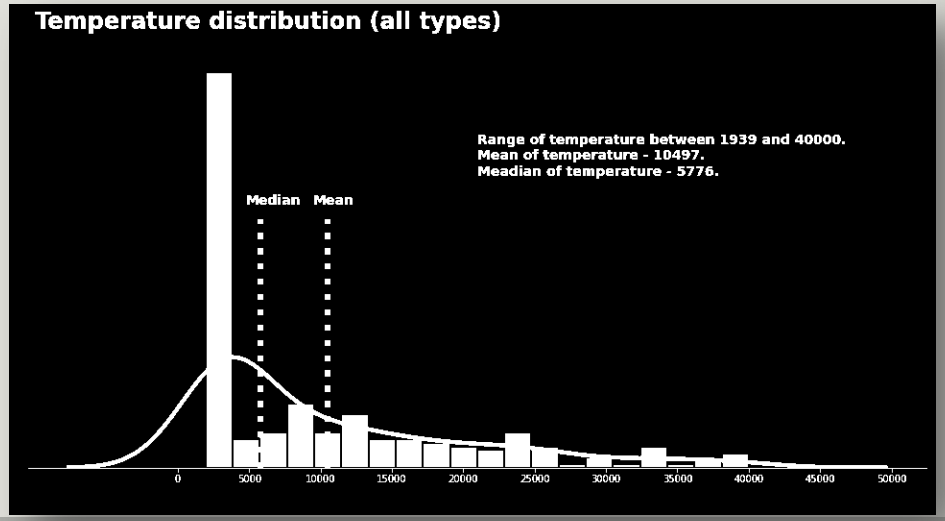



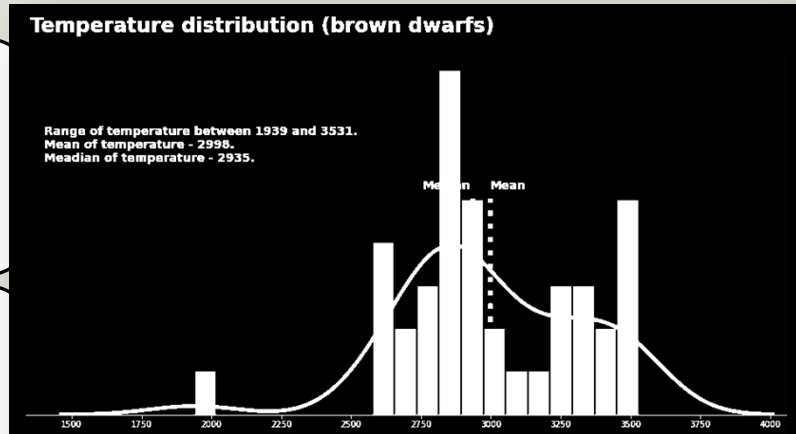
**3. Сделать выводы**

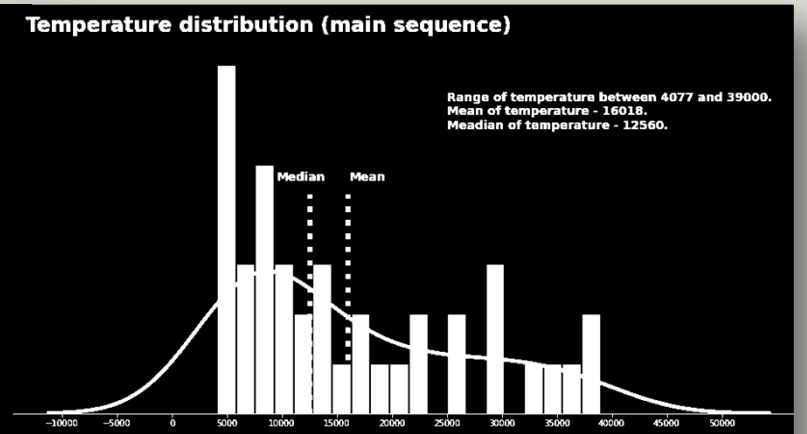
**Пример визуализации и выводов для одного из признаков - температуры:**

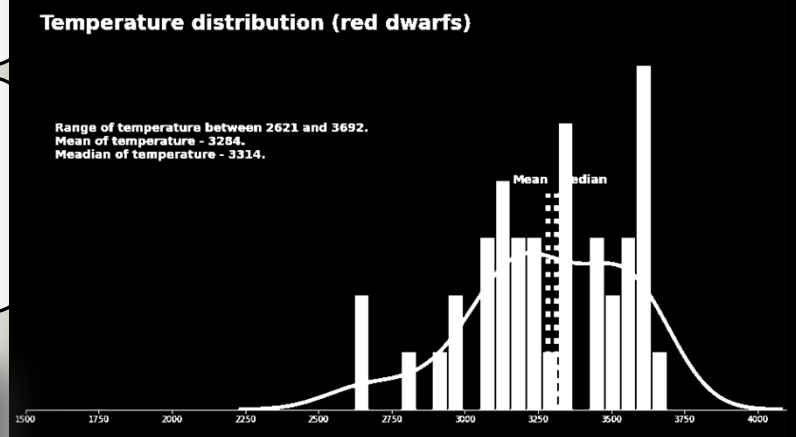
**1) распределение по температуре отличается от нормального;**

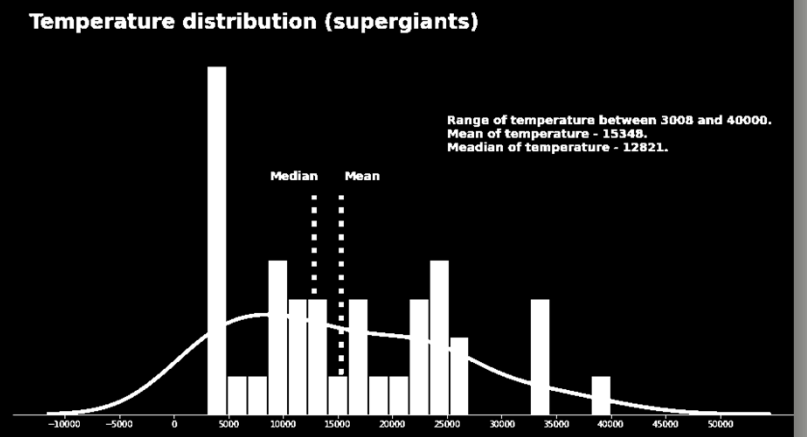
**2) наибольшую температуру имеют звезды главной последовательности, супергиганты и гипергиганты;**

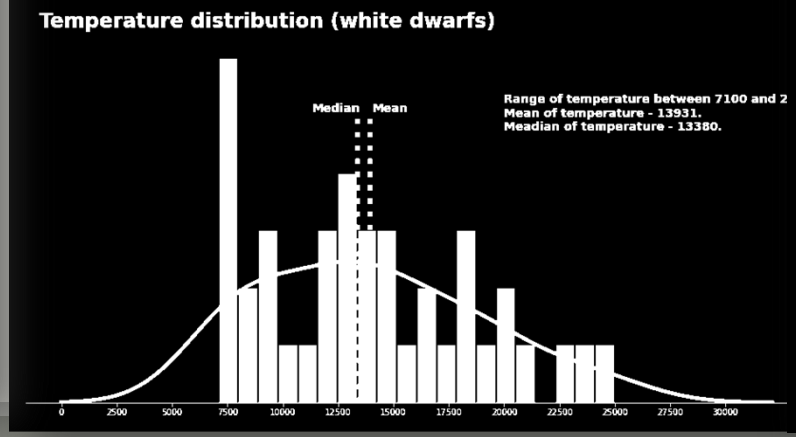


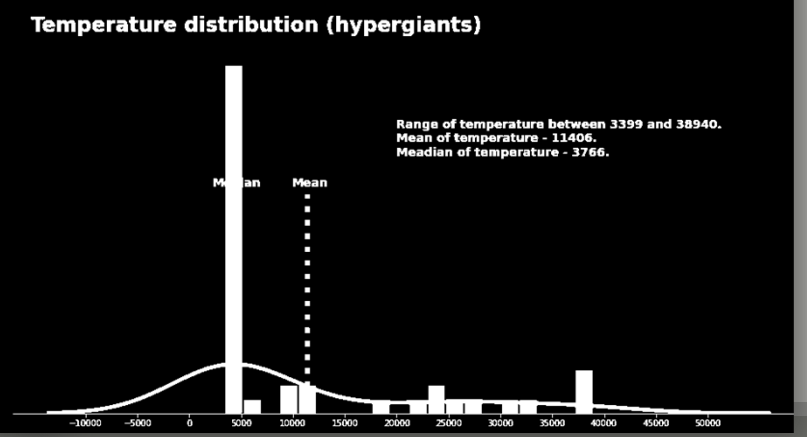



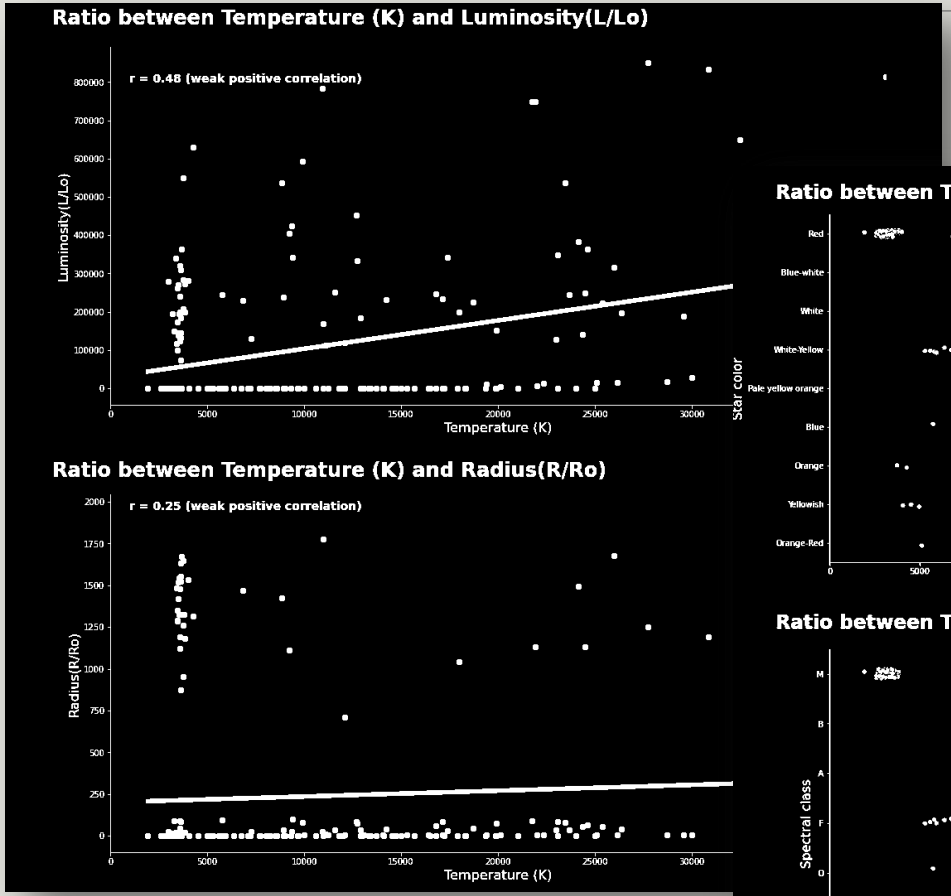



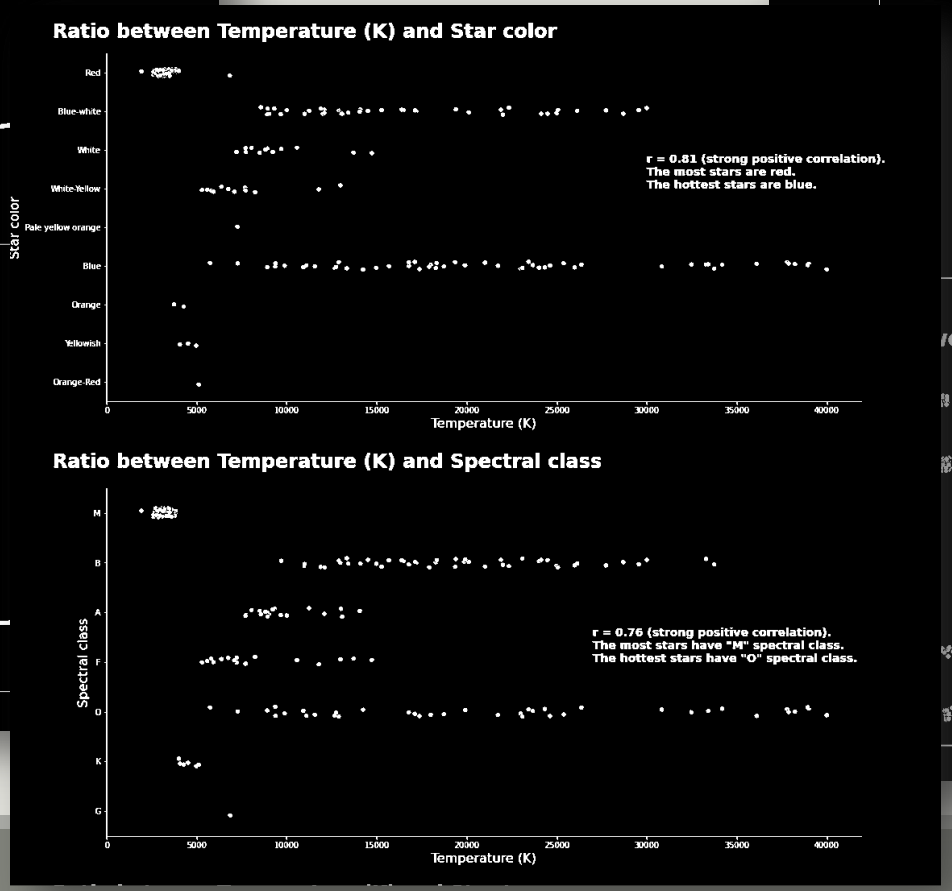



3) имеется сильная положительная корреляция температуры со спектральным  классом звезды и цветом;

4) корреляция температуры с целевой переменной - средняя.



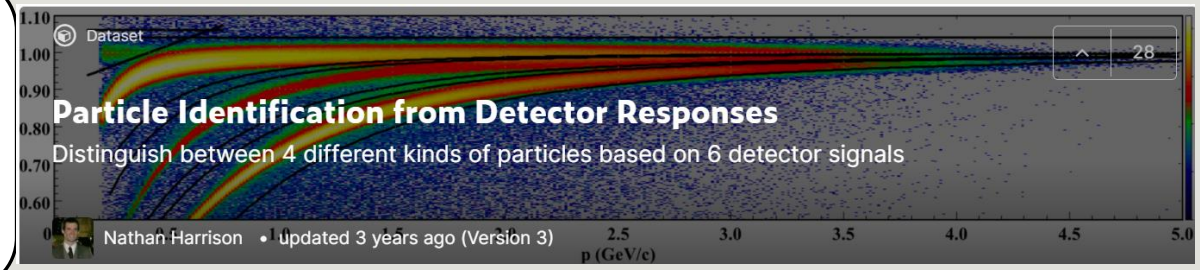
# Машинное обучение в астрономии и астрофизике: задачи, факты и датасеты

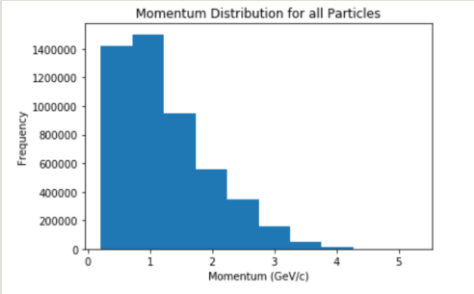
| | |
|---|---|
| **Распознавание экзопланет**<br>В 2017 году использование нейросети, созданной инженерами из Google Brain, привело к открытию двух новых экзопланет по данным с космического телескопа «Кеплер» | **Kepler Exoplanet Search Results**<br>https://www.kaggle.com/nasa/kepler-exoplanet-search-results |
| **Классификация объектов: звезда, галактика или квазар?**<br>Sloan Digital Sky Survey (SDSS, — «Слоуновский цифровой небесный обзор») — проект широкомасштабного исследования многоспектральных изображений и спектров красного смещения звёзд и галактик | **Sloan Digital Sky Survey DR14. Classification of Stars, Galaxies and Quasars**<br>https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey |
| **Классификация объектов: типы галактик**<br>В 2007 стартовал краудсорсинговый проект Galaxy Zoo. Пользователям выводится изображение объекта и нужно ответить на несколько вопросов о том, что они видят. Таким образом, набирается большая база размеченных данных, на которых можно обучать алгоритмы машинного обучения. | **Galaxy Zoo 2: Images. Shape Galaxy Clustering**<br>https://www.kaggle.com/jaimetrickz/galaxy-zoo-2-images |

# Машинное обучение в физике элементарных частиц: задачи, факты и датасеты

| | |
|---|---|
| **“Обучение обнаружению” (специфическая разновидность классификации)**<br>В 2014 году ЦЕРН провел конкурс на лучший алгоритм по поиску событий распада бозона Хиггса. Целевая функция, представляющая значимость открытия новой частицы, в работах 10 лучших участников составила 3.76 σ -3.80 σ, в то время как широко используемые модели без машинного обучения давали не более 3.50 σ.<br>Вдохновленные этими результатами, представители ЦЕРН организовали еще несколько соревнований на основе датасетов с реальными данными экспериментов. | Higgs Boson Machine Learning Challenge<br>https://www.kaggle.com/c/higgs-boson |
| | «Flavours of Physics: Finding τ → μμμ.<br>Identify a rare decay phenomenon»<br>**https://www.kaggle.com/c/flavours-of-physics** |
| Идентификация частиц по откликам в детекторе | Particle Identification from Detector Responses<br>**https://www.kaggle.com/naharrison/particle-identification-from-detector-responses/version/2** |

# Пример: разведочный анализ, добавление новых признаков, классификация

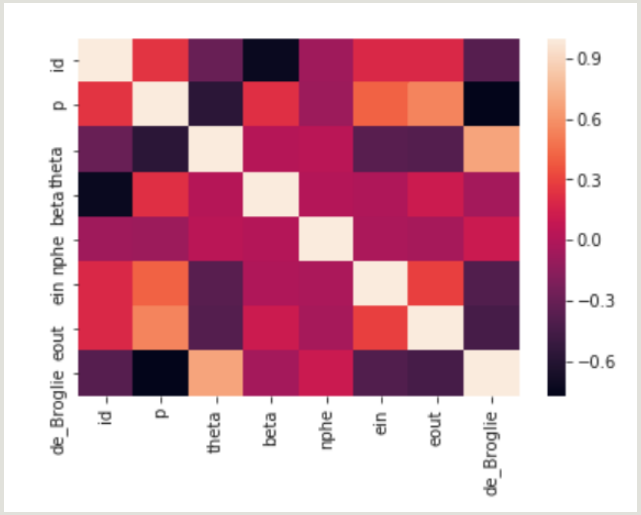



| | id | p | theta | beta | nphe | ein | eout |
|---|---|---|---|---|---|---|---|
| 0 | 211 | 0.780041 | 1.081480 | 0.989962 | 0 | 0.000000 | 0.000000 |
| 1 | 211 | 0.260929 | 0.778892 | 0.902450 | 0 | 0.000000 | 0.000000 |
| 2 | 2212 | 0.773022 | 0.185953 | 0.642428 | 4 | 0.101900 | 0.000000 |
| 3 | 211 | 0.476997 | 0.445561 | 0.951471 | 0 | 0.000000 | 0.000000 |
| 4 | 2212 | 2.123290 | 0.337332 | 0.908652 | 0 | 0.034379 | 0.049256 |

```
print(classification_report(newy_true, rfcpredict2))
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -11.0 | 0.98 | 0.96 | 0.97 | 2954 |
| 211.0 | 0.92 | 0.90 | 0.91 | 2996 |
| 321.0 | 0.91 | 0.93 | 0.92 | 2956 |
| 2212.0 | 0.97 | 0.98 | 0.98 | 2972 |
| micro avg | 0.94 | 0.94 | 0.94 | 11878 |
| macro avg | 0.94 | 0.94 | 0.94 | 11878 |
| weighted avg | 0.94 | 0.94 | 0.94 | 11878 |




positron (-11), pion (211), kaon (321), and proton (2212)

| | id | p | theta | beta | nphe | ein | eout | de_Broglie |
|---|---|---|---|---|---|---|---|---|
| 3139059 | 2212.0 | 2.577550 | 0.270532 | 0.943780 | 0.0 | 0.023163 | 0.123140 | 0.481015 |
| 1136605 | 211.0 | 0.232090 | 0.986443 | 0.834388 | 0.0 | 0.000000 | 0.000000 | 5.342066 |
| 4333018 | -11.0 | 0.334325 | 0.573388 | 1.010010 | 0.0 | 0.000000 | 0.000000 | 3.708487 |
| 4795395 | 321.0 | 2.487070 | 0.421448 | 0.979603 | 0.0 | 0.026642 | 0.245136 | 0.498514 |
| 3206407 | -11.0 | 0.967269 | 0.184006 | 0.999795 | 79.0 | 0.152941 | 0.064500 | 1.281794 |

# Заключение

Разработана система учебно-исследовательских заданий для занятий компьютерного практикума по вычислительной физике.

Показаны возможности использования наборов данных, размещенных в открытом доступе для ознакомления студентов с примерами решения фундаментальных проблем современной физики и астрофизики с помощью алгоритмов машинного обучения.

Работа с датасетами, рассмотренными в примерах, была апробирована в рамках курсового проектирования в 2020-2021 году. В 2021-22 годах практические работы на основе этих данных включены в программу практикума по дисциплине «Вычислительная физика».

# СПАСИБО ЗА ВНИМАНИЕ!
;)