Załącznik nr 4 do Decyzji nr 9/2019 Dziekana Wydziału EAIiIB z dnia 28 listopada 2019 r. Załącznik nr 4 do Zarządzenia Nr 14/2019 Rektora AGH z dnia 10 kwietnia 2019 r.

Kraków, dnia 17.05.2020r.

Bartłomiej Cerek
Imiona i nazwisko studenta

Prof. Adrian Horzyk
Imiona i nazwisko opiekuna pracy dyplomowej

Rozwój narzędzi do monitorowania jakości danych, dla eksperymentu ALICE w CERN, z użyciem metod uczenia maszynowego.
Tytuł pracy dyplomowej w języku polskim

Development of tools for data quality control for ALICE experiment at CERN, using machine learning methods.
Tytuł pracy dyplomowej w języku angielskim

STRESZCZENIE

Eksperyment ALICE w CERN wytwarza wielkie ilości danych, które są wykorzystywane do badań z zakresu fizyki jądrowej. Zanim naukowcy będą mogli podjąć się analiz, wszystkie anomalie pomiarowe muszą zostać wyeliminowane aby zapewnić wysoką jakość wyników. Ze względu na wielowymiarowość próbek, manualna detekcja odchyleń jak i użycie prostych parametrów statystycznych są trudne w przeprowadzeniu. W poniższej pracy zostały zaproponowane alternatywne rozwiązania problemu, bazujące na uczeniu maszynowym. Nadzorowane, nienadzorowane i częściowo nadzorowane przypadki zostały zaimplementowane, omówione i porównane. Rozważane są zarówno, klasyczne modele jak DBSCAN czy Isolation Forest, jak i techniki uczenia głębokiego (ang. deep learning) jak autoenkodery. Ponadto, przeprowadzona została dogłębna analiza przykładowego zbioru danych. Ponieważ budowanie zaufania użytkowników do algorytmów sztucznej inteligencji w dziedzinach badawczych jest kluczowe, problem wyjaśniania decyzji modeli został zaadresowany poprzez użycie techniki LIME. Wynikami projektu są pełne procedury wykrywania anomalii, które mogłyby zostać zaimplementowane w rzeczywistym środowisku w eksperymencie ALICE.

SUMMARY

The ALICE experiment at CERN is producing great amounts of data, that are being used to research nuclear physics. Before scientists can perform analysis, all measurements anomalies should be discarded to ensure high quality of results. Because of the high dimensionality of samples, it is difficult to perform outlier detection manually or using simple statistical parameters. In the following thesis, alternative, machine-learning-based solutions are proposed. Supervised, unsupervised, and semisupervised cases are implemented, discussed, and compared. Both, classic models like DBSCAN or Isolation Forest, and deep learning approaches like autoencoders, are taken into consideration. Beyond that, an in-depth analysis of the example dataset is performed. As building users' trust in artificial intelligence algorithms is crucial in research fields, the problem of model explainability is addressed by using the LIME framework. Outcomes of the project are full anomaly detection pipelines, that could be implemented in real environment ad ALICE experiment.