МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Факультет математики, фізики та інформаційних технологій

(повне найменування інституту, назва факультету (відділення))

Кафедра математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота

на здобуття ступеня вищої освіти «бакалавр»

(освітньо-кваліфікаційний рівень)

на тему «Аналіз винятків за допомогою алгоритмів машинного навчання в ElasticSearch»

«Analysis of faults behind the help of machine learning algorithms in ElasticSearch»

Виконав: студент денної форми навчання
спеціальності 126 – Інформаційні системи та технології .

(шифр і назва напряму підготовки, спеціальності)

Закап Назар Дмитрович

(прізвище, ім'я, по-батькові)

Керівник: кан.ф.-мат.н., доц. Петрушина Т.І.

(науковий ступінь, вчене звання, прізвище та ініціали, **підпис**)

Рецензент: Малахов Є.В

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:
Протокол засідання кафедри
№___ від «___»_____ 2022 р.

Завідувач кафедри
_____ Є.В. Малахов
(підпис) (прізвище, ініціали)

Захищено на засіданні ЕК № ___
протокол № __ від «___» ____ 2022 р.
Оцінка _____/_____/_____
(за національною шкалою, шкалою ECTS, бали)
Голова ЕК
_____ Н.Ф. Казакова
(підпис) (прізвище, ініціали)

Одеса – 2022

АННОТАЦІЯ

Метою дипломної роботи являється отримання якісних результатів інтелектуального аналізу на прикладі певного датасата за допомогою пошукового двигуну на основі повнотекстового пошуку Lucene Elastic Search.

Для візуалізації результатів використано компонент Kibana. який отримує дані від Elastic Search у вигляді JSON-запитів.

Результатом дипломної роботи являються моделі для задачі виявлення викидів, регресії та класифікації.

Робота кожної із моделей перевірена на датасеті, який представляє собою продаж розумних годинників.

ANNOTATION

The purpose of the thesis is to obtain qualitative results of data mining on the example of a certain date using a search engine based on full-text search Lucene Elastic Search.

The Kibana component was used to visualize the results. which receives data from Elastic Search in the form of JSON requests.

The result of the thesis work of the model for the task of identifying outliers, regression and classification.

The work of each of the models is tested on the dataset, which is the sale of smart watches.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ І ТЕРМІНІВ

ES – Elastic Search

API – Application Programming Interface

ELK – абревіатура для описання стеку: ElasticSearch, Logstash, Kibana

HTTP – HyperText Transfer Protocol

δ – параметр дельта

e – похибка

ВСТУП

В умовах розвитку інформаційних технологій одне із важливих місць займає інтелектуальний аналіз даних. На сьогоднішній день існує ряд задач, які успішно вирішуються методами інтелектуального аналізу. Велика кількість комерційних компаній проводять аналіз даних методами Data Mining та на основі результатів приймають рішення. В основі інтелектуального аналізу існує певна кількість математичних методів, таких як оптимізація, генетичні алгоритми, статистика, розпізнавання образів, прогнозування, виявлення викидів.

Для підготовки даних можна використовувати такий інструмент як ElasticSearch. ES створювався як двигун для пошуку даних та для індексації документів, однією із його переваг являється нечіткий пошук, за допомогою якого можна знайти всі відповідності в системі. В системі дані зберігаються у вигляді індексів, які організовані у форматі JSON[3].

З виходом нових оновлень в ES стало можливим розв’язати наступні задачі інтелектуального аналізу:

- виявлення відхилень та викидів;
- прогнозування подій та процесів або регресія;
- віднесення об’єктів до певної категорії або класифікація.

Ціль роботи заключається в отриманні якісних результатів аналізу даних при використанні порівняно простого для використання інструмента ES с традиційними методами інтелектуального аналізу.

Засоби інтелектуального аналізу широко використовуються в різних сферах діяльності. Для цього можна використовувати ES, так як даний кластер являється ефективним засобом для аналізу даних за рахунок масштабування, високої швидкості роботи та легкому управлінні.

Практична цінність роботи заключається в отриманні об’єктивних результатів, які є основою для прийняття рішень.

Метою дипломної роботи являється обробка даних інструментом ES для отримання якісних результатів та створення моделі для різних задач інтелектуального аналізу. Для того, щоб досягнути даної цілі необхідно вирішити наступні задачі:

- провести порівняльний аналіз існуючих інструментів інтелектуального аналізу даних;
- підготовити відповідним образом набір даних;
- створити модель ES та виконати конфігурацію даної моделі;
- виконати аналіз даних для задач класифікації, регресії та виявлення викидів на справжньому датасеті;
- запропонувати наглядну візуалізацію за допомогою інструмента Kibana;

Об’єктом роботи являється модель аналізу на основі підходу інструмента ES.

Предметом дослідження являються методи аналізу даних на основі інструмента ES.

ВИСНОВОК

У процесі виконання дипломної роботи отримано якісні результати аналізу за допомогою пошукового двигуну ES. Для вирішення задач виявлення викидів, регресії та класифікації створені відповідні моделі. Навчання та тренування кожної із моделей було продемонстровано на конкретному датасеті розумних годинників.

Для кожної із задач була поставлена ціль і методи, які дозволяють досягнути даної цілі.

Для задачі виявлення викидів визначено поля, які являються ознаками поведінки моделі та алгоритми з вказанням гіперпараметрів.

Для задачі регресії визначені незалежні змінні та одна залежна змінна з вказанням гіперпараметрів.

Для задачі класифікації визначено залежне поле та незалежні поля з вказанням гіперпараметрів.

Результати візуалізовано за допомогою інструмента Kibana для кращої інтерпретації результатів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. ELK Stack Tutorial: What is Kibana, Logstash & Elasticsearch [Електронний ресурс] – https://www.guru99.com/elk-stack-tutorial.html
2. Почему Elasticsearch — хороший выбор для сбора и анализа данных среднего объёма [Електронний ресурс] – https://tproger.ru/blogs/why-elasticsearch-is-a-good-choice/
3. Метод ближайших соседей [Електронний ресурс] – http://surl.li/cdugv
4. Классификация данных методом k-ближайших соседей [Електронний ресурс] – https://loginom.ru/blog/knn
5. How Filebeat works [Електронний ресурс] – https://www.elastic.co/guide/en/beats/filebeat/current/how-filebeat-works.html
6. Интеллектуальный анализ данных (Intelligent data analysis) [Електронний ресурс] – https://wiki.loginom.ru/articles/data-analysis.html
7. Метод k-ближайших соседей (k-nearest neighbour) [Електронний ресурс] – https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19
8. Введение в Elastic Stack - что такое и где использовать? [Електронний ресурс] – https://oracle-patches.com/is/введение-в-elastic-stack
9. Методы регрессионного анализа в Data Science [Електронний ресурс] – https://habr.com/ru/company/otus/blog/485972/
10. Основы линейной регрессии [Електронний ресурс] – http://surl.li/cdtwg
11. Коэффициент детерминации [Електронний ресурс] – http://statistica.ru/theory/koeffitsient-determinatsii-i-lineynaya-regressiya/
12. Обнаружение выбросов с помощью локального коэффициента выброса (LOF) [Електронний ресурс] –

https://runebook.dev/ru/docs/scikit_learn/auto_examples/neighbors/plot_lof_outlier_detection

13. Назначение и основные функциональные возможности [Електронний ресурс] – https://www.bigdataschool.ru/wiki/elasticsearch
14. Accelerate development and improve your application code [Електронний ресурс] – https://www.elastic.co/observability/application-performance-monitoring
15. Функция потерь Хьюбера [Електронний ресурс] – https://gnezdoparanoika.ru/stati/6458-funkciya-poter-hyubera.html
16. Fleet and Elastic Agent [Електронний ресурс] – https://www.elastic.co/guide/en/fleet/current/fleet-overview.html
17. Fleet in Kibana enables you to manage Elastic Agent [Електронний ресурс] – https://www.elastic.co/guide/en/kibana/current/fleet.html
18. Классификация и прогнозирование в интеллектуальном анализе данных [Електронний ресурс] – https://russianblogs.com/article/5317968932/
19. Типы аномалий в видеоизображениях [Електронний ресурс] – https://cyberleninka.ru/article/n/tipy-anomaliy-v-videoizobrazheniyah
20. What is an Elasticsearch Index? [Електронний ресурс] – https://www.elastic.co/blog/what-is-an-elasticsearch-index
21. Организация сбора и парсинга логов при помощи Filebeat [Електронний ресурс] – https://habr.com/ru/post/550352/
22. Новые возможности анализа табличных данных с алгоритмами машинного обучения в Elastic [Електронний ресурс] – https://habr.com/ru/company/step_logic/blog/546246/