MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susţinere
Şef de departament:
Fiodorov I. dr., conf.univ.

„___” _____ 2022

Analiza exploratorie a datelor – capcane și piste false

Teză de master

Student: _____ **Fridman Stanislav, TI-201M**

Coordonator: _____ **Prof. dr Leahu Alexei**

Consultant: _____ **Lector. univ. mag. Cojocaru Svetlana**

Chişinău, 2022

АННОТАЦИЯ

Дипломный проект представляет собой исследование капканов в процессе сбора и разведочного анализа данных. Актуальность темы заключается в том, чтобы отличать правду от лжи, когда идет речь о сборе данных, статистике, популяции, взятых выборках. Также раскрываются вопросы формирования и манипулирования общественным мнением. Целью работы является поиск и выявление тех капканов и также работа с вычислительными инструментами, беря в учет найденные капканы.

В работе проведен анализ действительных методов разведочного анализа данных, идентифицированы капканы и ложные пути, преимущества и недостатки и области применения каждого метода. Аналогично проведен обзор методов сбора данных, выявлены капканы и установлена связь между капканами сбора данных и их анализа. Для практического решения задачи поиска капканов выбрана актуальная в наши дни тема, связанная с пандемией COVID-19. Использованы вычислительные инструменты языка Python, предназначенные для исследования статистики и поиска возможной неправды – pandas и matplotlib. Работа включает в себя знания из курсов разведочного анализа данных и вычислительной статистики.

REZUMAT

Proiect de master este studiul capcanelor în procesul de colectare și analiză exploratorie a datelor. Relevanța subiectului este de a distinge adevărul de minciună atunci când vine vorba de colectarea datelor, statistici, populație, eșantioane prelevate. Sunt dezvăluite și problemele formării și manipulării opiniei publice. Scopul lucrării este găsirea și identificarea acelor capcane și, de asemenea, lucrul cu instrumente de calcul, ținând cont de capcanele găsite.

Lucrarea analizează metode valide de analiză exploratorie a datelor, identifică capcane și piste false, avantaje și dezavantaje și domenii de aplicare ale fiecărei metode. În mod similar, metodele de colectare a datelor au fost revizuite, au fost identificate capcane și a fost stabilită o legătură între colectarea datelor și capcanele de analiză. Pentru o soluție practică a problemei găsirii capcanelor, a fost selectat un subiect relevant legat de pandemia COVID-19. Au fost folosite instrumentele de calcul ale limbajului Python, concepute pentru a studia statisticile și a găsi posibile falsuri - pandas și matplotlib. Lucrarea include cunoștințe de la cursuri de analiză exploratorie a datelor și statistică computațională.

ABSTRACT

This master degree project is the study of traps in the process of collecting and exploratory data analysis. The relevance of the topic is to distinguish truth from lies when it comes to data collection, statistics, population, samples taken. The issues of the formation and manipulation of public opinion are also disclosed. The purpose of the work is to find and identify those traps and also work with computing tools, taking into account the traps found.

This project analyzes valid methods of exploratory data analysis, identifies traps and false ways, advantages and disadvantages and areas of application of each method. Similarly, data collection methods were reviewed, traps were identified, and a link between data collection and analysis traps was established. Relevant topic related to the COVID-19 pandemic has been selected for a practical solution to the problem of finding traps. The computational tools of the Python language were used, designed to study statistics and find possible falsehoods - pandas and matplotlib. This project includes knowledge from courses in exploratory data analysis and computational statistics.

Содержание

ВВЕДЕНИЕ

Разведочный анализ данных – предварительное исследование основных свойств данных, поиск общих закономерностей, характера, свойств, различий и особенностей в заданных или случайных массивах данных. Целью разведочного анализа данных является нахождение связи между переменными при отсутствии информации, характеризующей их происхождение.

Существуют общепринятые практики и математические модели ведения анализа данных – анализ распределений переменных, корреляционный анализ, факторный анализ, дискриминантный анализ, многомерное шкалирование, анализ гистограмм. Однако, можно использовать эти методы в целях обмана аудитории и манипулирования общественным мнением. Результаты подобного исследования могут влиять на выбор определенного продукта или кандидатуры.

Целью этой работы является исследование возможных случайных или умышленных капканов и ложных путей в разведочном анализе данных, а также поиск решения для автоматизации расчетов, имея в виду исследованные капканы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ:

1. Корреляционный анализ [Электронный ресурс]. - Режим доступа: https://www.statmethods.ru/statistics-metody/korrelyatsionnyj-analiz/

2. Многомерное шкалирование [Электронный ресурс]. - Режим доступа: https://www.statmethods.ru/statistics-metody/mnogomernoe-shkalirovanie/

3. Почему с нормальным распределением не все нормально [Электронный ресурс]. - Режим доступа: https://habr.com/ru/post/191438/

4. Типичные регрессионные ошибки [Электронный ресурс]. - Режим доступа: https://pokrovka11.wordpress.com/2017/11/27/типичные-регрессионные-ошибки

5. Leahu A. Analiza exploratorie a datelor [Текст] / Leahu A. – Chisinau, 2019 – 78 с.

6. Дембицкий С. Выборка, сбор и анализ данных в исследованиях смешанного типа [Электронный ресурс]. - Режим доступа: http://soc-research.info/mixed/5.html

7. Ошибки анкетных опросов. 2 ошибка: формулировка анкеты. 13 случаев непонимания и манипуляций в опросе (1 часть) [Электронный ресурс]. - Режим доступа: https://habr.com/ru/post/308938/

8. COVID-19: Five Common Statistics Errors – and How to Avoid Them - The Wire Science [Электронный ресурс]. – Режим доступа: https://science.thewire.in/health/covid-19-five-common-statistics-errors-and-how-to-avoid-them/э

9. Находим аномалии в российской статистике COVID-19 / Хабр [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/587596/

10. Моя «корона». Как в поликлинике добиться теста на COVID-19 и получить лечение [Электронный ресурс]. – Режим доступа: https://fn-volga.ru/news/view/id/146915

11. «Тесты у всех подряд не берем»: семье отказывают в диагностике COVID-19 [Электронный ресурс]. – Режим доступа: https://sibkray.ru/news/1/939215/

12. Exploring Casual COVID-19 Data Visualizations on Twitter: Topics and Challenges [Электронный ресурс]. – Режим доступа: https://www.mdpi.com/2227-9709/7/3/35

13. Всемирный информационный ресурс [Электронный ресурс]. - Режим доступа: https://www.google.com