http://aptitude.w3c.fmi.uni-sofia.bg/

Н А У Ч Е Н О Т Ч Е Т

По проект APTITUDE

Иновативна софтуерна платформа за анализи на големи масиви от учебни и игрови данни за ориентирана към потребителя адаптация на технологично подпомогнато обучение

Фонд „Научни изследвания“, МОН,

Конкурс за финансиране на фундаментални научни изследвания по обществени предизвикателства – 2018 г.

Номер на договор: КП-06-ОПР03/1 от 13.12.2018г.

D3.1. Моделиране на данни от модули от учебни курсове от LMS и образователни игри

ФОНД НАУЧНИ ИЗСЛЕДВАНИЯ
МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА

I. Цел на документа

Основната цел на документа е да се предложи процедура за моделиране на данни от модули от учебни курсове от системи за управление на обучението (LMS) и образователни игри. Процедурата включва четири основни етапа, а именно:

1) Подготовка на данни – включва предложения модел на анонимизация на данни;
2) Изчистване на данни;
3) Подбор на данни за обработка;
4) Предварителна обработка на данни за обучение и валидиране на набор данни.

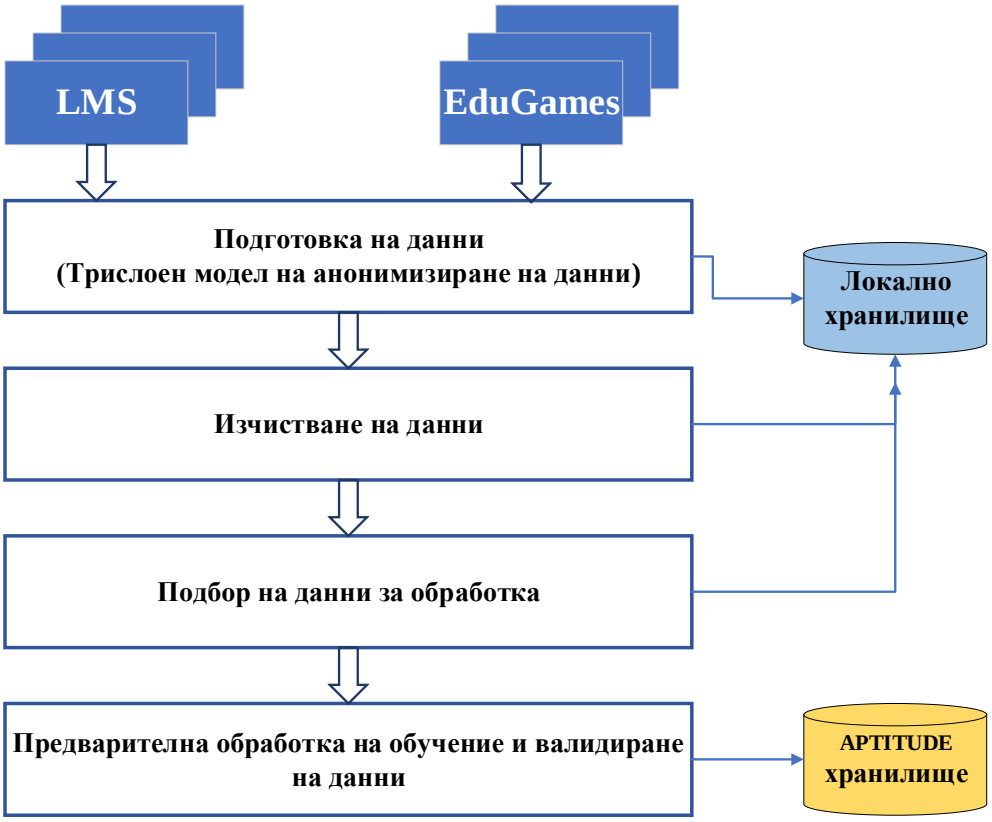
В резултат данните се съхраняват в изграденото хранилище (D2.3), където се интегрират данни от различни източници и се използва формата на данни описан в спецификациите (D2.1 и D2.2).

II. Обща процедура за моделиране на данни от модули от учебни курсове от LMS и образователни игри.

Процедурата за моделиране на данни от модули от учебни курсове от LMS и образователни игри следват техниките за обработка за откриване на полезни знания от колекция от данни, които обхващат следното: подготовка, изчистване и подбор на данни; откриване на знания и вземане на решения, включващи резултати и интерпретиране на точни решения от наблюдаваните резултати.

Задълженията на фазата на предварителна обработка е предварителна обработка на обучението и валидиране на набори от данни. Предварителната обработка на данните при откриване на знания обхваща изчистване на данните по отношение на точността и избор на функции по отношение на уместност и извличане на функции. Целта е да се създаде модел на хранилище, като се използват набори от данни за обучение и валидиране и да се прилагат различни алгоритми за машинно обучение като класификация и клъстери.

Процедурата за моделиране на данни от модули от учебни курсове от LMS и образователни игри за прогнозиране на обучението на студентите, базиран на машинно обучение, е представен на фиг. 1.



Фиг. 1. Обща процедура за моделиране на данни от LMS и образователни игри.

Различните LMS и образователните игри имат различна структура и начина на съхраняване на данни, които се отнасят за учебният процес. Всяка една система/игра предоставя собствено API, което да обменя данни с други системи.

След извличането на данните от съответния източника следват 4 етапа за моделиране на данни, които са:

1) Подготовка на данни;
2) Изчистване на данни;
3) Подбор на данни за обработка;
4) Предварителна обработка на данни за обучение и валидиране на набор данни.

1. Подготовка на данни

Основният процес в този етап е анонимизация на данни. Анонимността на данни е процесът на премахване на лична информация от данните. Прави се по такъв ред, за да се гарантира запазването на личния живот на хората. С други думи, субектът на данни вече не може да бъде идентифициран. В контекста на медицинските данни анонимните данни се отнасят до данни, от които пациентът не може да бъде идентифициран от получателя на

информацията. Името, адресът и пълният пощенски код трябва да бъдат премахнати, заедно с всяка друга информация, която във връзка с други данни, които се съхраняват или разкриват на получателя, биха могли да идентифицират пациента.

Деанонимността е обратният процес, при който анонимните данни се препращат с други източници на данни за повторно идентифициране на анонимния източник на данни. Например, данните от преброяването могат да бъдат публикувани за статистически цели, но публичното оповестяване с всички имена, адреси, пощенски кодове и други идентифицируеми данни се премахва.

Няколко общи типове анонимизация на данни са следните:

Отстраняване

Напълно премахване на полета, които биха могли да бъдат използвани по всякакъв начин за идентифициране на човек. Счита се за силна форма на анонимизация на данни.

Модификация

Реакцията включва премахване и други техники, като изтласкване на данни на хартия с маркер и правене на фотокопие на резултата

Шифроване

Шифроването може да бъде много силно и трудно да се обърне обратно. То представя няколко предизвикателства като генерирането на достатъчно силен ключ за декриптиране. Анонимизацията на данни не е предназначена да бъде обратима, така че управлението на ключовете за декриптиране също е проблем. В идеалния случай би бил генериран силен и напълно случаен ключ и след това незабавно изтрит от паметта, когато завърши шифроването.

Хеширане

Този метод е по-добър от Шифроването, тъй като няма ключова двойка, която да криптира / дешифрира данните. Използването на силни и надеждни хеш-алгоритми в едно направление е най-вероятно най-добрият начин за постигане на желано ниво на анонимност на личната информация.

Ние предлагаме модификация на хешированието, наречена Squeezed Hashing. За да бъде теоретично и практически невъзможно възстановяването на оригиналните данни, някои от битовете от генерираното съобщение могат да бъдат пропуснати - например първите 4 еднакви бита и последните 4 еднакви бита от дайджеста могат да бъдат изтрити (пропуснати). При този начин на генериране дайджестът на съобщенията ще бъде еднакъв за едни и същи входни данни. И също така - за целите на анализа и отчитането устойчивостта на сблъсък на хеш алгоритъма не е важна (ако изобщо има такава).

Например, типичното име на обучаемия (първо, второ и фамилно име) съдържа между 6-15 знака всеки (с други думи - 6-15 байта). Ако се използва хеш функция, генерираща дайджест на съобщения с повече от 15 байта, се използва пропускането на 8 бита (четири в началото и четири в края на дайджеста) ще запази до известна степен уникалността на картографирането. В случай на прекъсване на използвания хеш алгоритъм, истинското име може да бъде намерено с груба сила в рамките на 28 = 256 повторения. Едно от възможното

подобрение е да се приложи друг алгоритъм за хеширане върху новосъздадения дайджест (дайджест на стиснатото хеш съобщение).

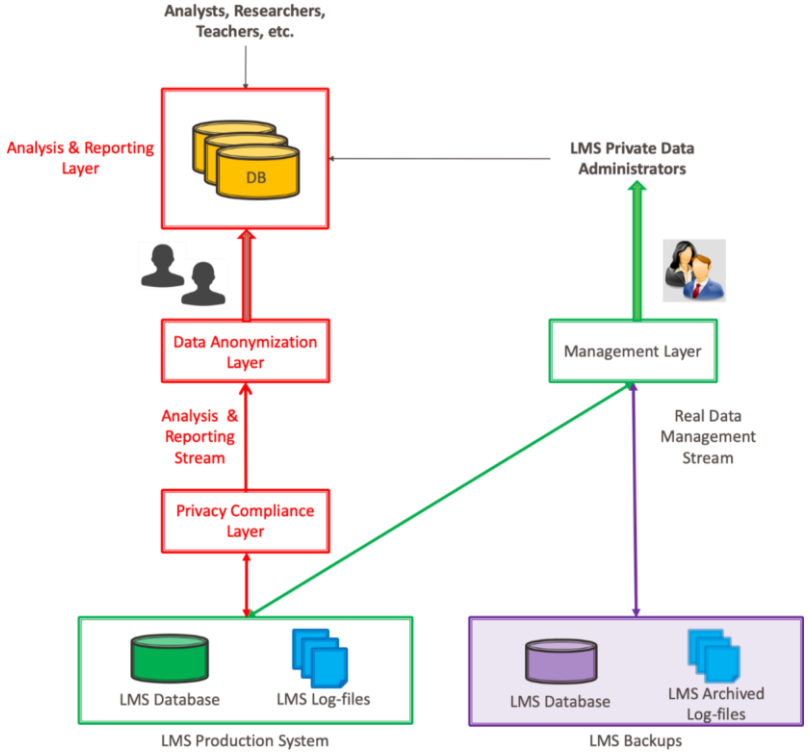
Добри представители на хеш функциите са SHA-2 (SHA-224, SHA-256, SHA-384, SHA-512) и SHA-3 (SHA3-224, SHA3-256, SHA3-384, SHA3-512) семейства - те произвеждат съобщения дайджести с дължина 28/32/48 / 64байта.

Маскиране на данни

Маскирането на данни е потенциално слаба форма на анонимизация на данни, която може да включва кодиране на данни и подмяна на символи. Предимството на маскирането на данни е, че то поддържа структурата на данните, така че числата остават числа, а датите остават дати. Това позволява анонимните данни да се използват за тестване на системата, без да се задействат грешки в приложението.

Различните LMS системи използват различна информация, схеми на база данни, политики за архивиране / възстановяване и др. Събирането на данни за целите на анализа и отчитането може да бъде различно от едната страна. И от друга, компаниите в ЕС, които управляват лична информация, са задължени да спазват разпоредбите на GDPR. Следователно, дори за целите на анализа и отчитането и статистиката, оригиналните данни не могат да бъдат използвани - или в известна степен.

Предлага се модел за анонимна поверителност на LMS (LMS Anonymized Privacy Model (APM)) е независим от всяка LMS система. Състои се от три слоя - интегриращ слой (наречен слой за спазване на поверителността или Privacy Compliance Layer), слой за анонимност на данни (Data Anonymization Layer) и слой за анализ и отчитане (Analysis & Reporting Layer) - фиг. 2.



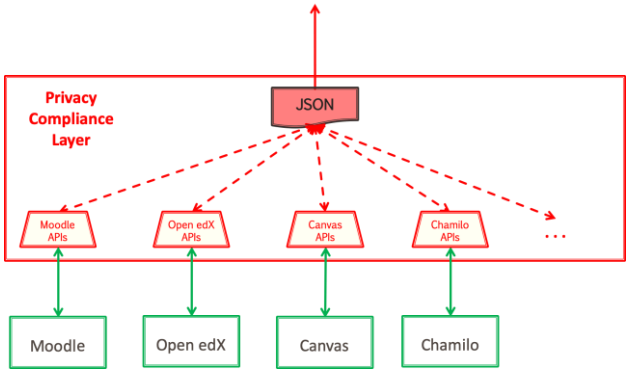
Фигура 2: Трислоен модел за анонимност на данни

Структурата на модела на данни се основава на трислоен модел за анонимност:

- Layer 1 - слой за спазване на поверителността - Privacy Compliance Layer;
- Layer 2 - слой за анонимност на данни - Data Anonymization Layer;
- Layer 3 - слой за анализ и отчитане - Analysis and Reporting Layer.

1.1. Слой 1 е слоят за спазване на поверителността „Privacy Compliance Layer“

Той съдържа политики, правила и процедури. Слоят за спазване на поверителността е рамка. Тя включва интеграции с различни LMS системи. Тази рамка е шлюз на API към всички слоеве над и под него. Той съдържа всякаква специфична информация (според LMS система - като схема на база данни, логическа структура на данните и т.н.), необходима за събиране на информация от конкретна LMS. И също така, той съдържа функции на API, за да може да предава данните в следващия слой - слой за анонимизация на данни (Data Anonymization Layer). Идеята е да има един универсален слой. Получените / събраните данни ще се съхраняват в локалната база данни на слоя за анализ и отчитане (Analysis and Reporting Layer). Този слой трябва да предаде събраните данни от LMS системите, заедно с информация за какъв тип данни се извличат (файл JSON). Фиг. 3 показва вътрешната архитектура на този слой.



Фигура 3: Слоят „Спазване на поверителността“ (Privacy Compliance Layer)

Правоъгълниците със стрелките в долната част представляват различни LMS системи (показани са само няколко).

За всяка LMS система е необходим специфичен модул „API“. Този модул API ще трябва да разкрие набор от функции на API (специфични за различните LMS), за да може да извлича информация от LMS.

След като информацията бъде извлечена от системата LMS, тя ще се трансформира във файл във формат JSON. Файлът JSON има четири атрибута: поле за данни (съдържа стойността на оригиналния атрибут), поле за име (името на оригиналния атрибут), тип поле (цяло число, низ, blob и т.н.), секретно поле / Secret Field (информацията е секретна или не е тайна) - в случай на „secret“ поле/полета-> секцията с данни трябва да бъде анонимна.

Следващият пример показва структура на файлов формат JSON, която съдържа цялата необходима информация за слоя „Анонимност на данни“ за обработка на данните. Той има името на полето, съдържа информация за подаденото - да бъде анонимно или не, съдържа информация за типа (за определяне на метода за анонимизация) и съдържа реалната предавана стойност:

```
[
 {
  "Name":          "e-mail",
  "Secret":        "sqhash",
  "Type":          "string",
  "Value":         "first.second@domain.com",
 },
 {
  "Name":          "First.Name.Of.Student",
  "Secret":        "hash",
  "Type":          "string",
  "Values":        "Ivaylo"
 },
 {
  "Name":          "Last.Name.Of.Student",
  "Secret":        "sqhash",
  "Type":          "string",
  "Values":        "Chenchev"
 },
 {
  "Name":          "Student.Age",
  "Secret":        "mask",
  "Type":          "integer",
  "Values":        24
 },
 {
  "Name":          "City",
  "Secret":        "false",
  "Type":          "string",
    "Values":      "Sofia"
 },
]
```

Полето „Secret“ съдържа следните стойности: [“hash” | “sqhash” | “masking” | “false”]. .

Ако е зададено на "hash" - тогава съдържанието на атрибута "Стойност" ще бъде анонимно с еднопосочна хеш функция.

Ако е зададено на „sqhash“ - тогава съдържанието на атрибута „Value“ ще бъде анонимно с еднопосочна хеш функция и 8 бита ще бъдат премахнати (както е предложено в предишния раздел с анонимизирането на „squeezed hash“). В горния пример Фамилното име на ученика ще бъде анонимно с метода „squeezed hash“.

Ако е зададено на „mask“ - тогава съдържанието на атрибута „Value“ ще бъде анонимно с маскиране (числото ще остане като число).

Ако полето „Secret“ е зададено на „false“ - тогава атрибутът „Value“ няма да бъде променен.

В примера, елементът Name = “City” няма да бъде анонимиран, тъй като не съдържа никаква лична информация.

Целта на анонимизацията не е да се изтрие или скрие цялата информация, а само личната информация. Разбира се, събраните в LMS системи данни трябва да се използват за анализ, статистика и отчитане. Следователно в примера Градът има нужда от информация, защото тя определя физическото местоположение.

Горният пример за предложен формат на JSON файл може да бъде адаптиран (разширен) за всеки конкретен случай с LMS система.

1.2. Слой 2 е слоят „Анонимност на данни“ или Data Anonymization Layer

Това е основният слой от нашия тристепенен модел на слой. Тук събраната информация от предишния слой ще бъде анонимна. Този процес се основава на получените данни и неговия тип (получен от предишен слой). Важно е да се гарантира картографирането на входните данни и генерираните „анонимни“ данни.

Един възможен начин е запазването на тези карти да се съхранява в локална база данни. Тази база данни ще има една цел - да запази предадените (оригиналните) стойности и „картографираните в“ стойности. Това изобщо не е желан метод, защото трябва да се приложи допълнителна сигурност за запазване на личната информация. Той също трябва да бъде съобразен с разпоредбите и стандартите за сигурност (като GDPR). Изглежда удвояване на информацията от оригиналната (оригиналните) система (системи) на LMS.

В най-добрия случай няма база данни, но силни и надеждни еднопосочни хеш-функции, които недвусмислено картографират входните данни в съобщенията за изваждане на съобщения - от едни и същи входни данни ще се генерира същият хеш-изход.

Както беше описано в предишния слой, различните видове входни данни ще трябва да бъдат управлявани по различен начин. Само секретната информация (като лично идентифицираната информация (имена, имейли)) ще бъде хеширана / обработвана.

За целите на анализа не е важна реалната възраст на учащия, а повече възрастовият диапазон. Фиг. 4 показва примера как може да бъде приложеното маскиране.

Age (10;15] -> to be replaced with “15”
Age (15;20] -> to be replaced with “20”
Age (20;25] -> to be replaced with “25”
…
Age (55;60] -> to be replaced with “60”
Age (60;99] -> to be replaced with “65”

Фигура 4: Маскиране на данни

След като входните данни се обработват, генерираният изход отново е JSON файл със сходна структура. Този JSON файл ще бъде прехвърлен към следващия слой за неговото съхранение и за по-нататъшна обработка.


```
[
 {
  "Name":                  "e-mail",
  "Secret":                "sqhash",
  "Type":                  "string",
  "Value":                 "first.second@domain.com",
 },
 {
  "Name":                  "First.Name.Of.Student",
  "Secret":                "hash",
  "Type":                  "string",
  "Values":
"2261ace9b7cdaa96d4980f9b08290f70de96ad769cc0677d8762208eaae469e8"
 },
 {
     "Name":                  "Last.Name.Of.Student",
  "Secret":  "sqhash",    "Type":  "string",
  "Values":
"366904cbb353151b64115f618859e5fdc81f166d23af3946ae27c948efb9fd09"   },
 {
  "Name":                          "Student.Age",
  "Type":                          "integer",
  "Values":                        25
 },
 {


  "Name":                          "City",
  "Type":                          "string",
```

```
    "Values":                    "Sofia"
  },
]
```

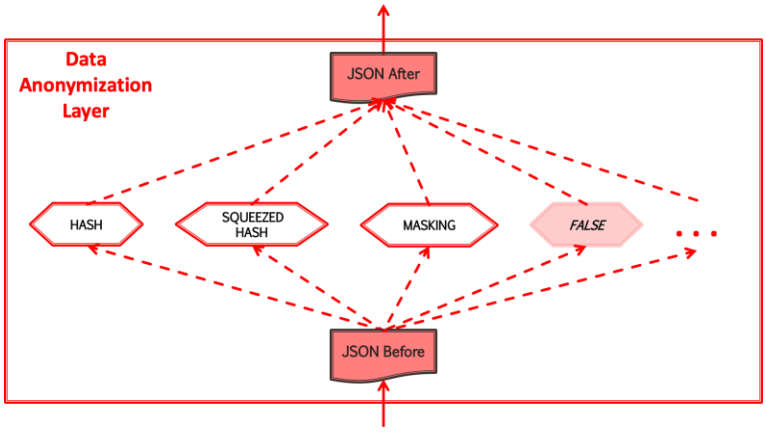
В примера JSON изход по-горе:

- използваната хеш функция за името на първия ученик е SHA3-256;

- използваната функция „sqhash" за името на последния ученик е SHA-256 и първите 4 еднакви бита, а последните 4 еднакви бита са премахнати, всички останали битове са изместени вляво; след това SHA3-256 се прилага върху новогенерираното съобщение (намаленото с 1 байт дайджест съобщение):

SHA3-256 (Squeeze8bits_and_Left_Shift (SHA-256 („Value"")))

На фиг. 5 се показва потока от данни в този слой „Анонимност на данни"(Data Anonymization Layer):



Фигура 5: Слоят „Анонимност на данни" (Data Anonymization Layer)

С други думи, този слой осигурява алгоритмичните механизми (инструментите), които ще се прилагат върху входните данни (получени от предишния слой).

2. Изчистване на данни

Вторият етап от предложената процедура е изчистването на данни. По време на тази дейност е наложително да се определи как данните могат да бъдат използвани за изграждане на желания модел и решаване на конкретния проблем. Има няколко основни типа проблеми, които ще бъдат описани в следващите редове.

A. Проблем 1: Липсващи и / или непълни данни, както и коригирането им

Доста често срещано в реалните данни е наличието на непълни записи [1]. Едно или повече от тези полета нямат стойност. Тази липса на пълнота на данните може да доведе до забавяне на обучението и до неточност в резултатите. Поради честото присъствие на този проблем са разработени различни начини за преодоляването му, в зависимост от данните и модела, който се изгражда.

B. Проблем 2: Данни с различен мащаб и обхват и тяхното нормализиране

При обучение на невронни мрежи, ако данните са с различни мащаби, това би довело до нестабилност и неточност в модела [2]. За да се избегне доминирането на една независима променлива над друга, самите данни трябва да бъдат мащабирани, или нормализирани. Този процес се нарича още нормализиране.

C. Проблем 3: Повредени, конфликтни и подвеждащи данни

Събирането на данни често е причинено от човешки или машинни грешки. В резултат на това данните не могат да бъдат интерпретирани погрешно от компютрите и това ги прави неизползваеми. За справяне с този проблем често се използва математическа регресия или групиране.

D. Проблем 4: Категорични данни

Това са данни, чиято стойност е в текстов формат, а не в цифров формат. Повечето алгоритми за машинно обучение използват само числа като вход и изход. Това се дължи на факта, че те основно използват математически функции и извършват сложни математически изчисления.

Основните дейности, които ще се предприемат на този етап е да се премахнат липсващи и/или непълни данни и да се групират данните по предварително избрани признаци.

3. Подбор на данни за обработка

След като входящите данни бъдат обработени, трябва да се реши кои променливи да се използват в модела. Изборът на правилните входни променливи е в основата на моделите за прогнозиране. През последните години, с натрупването на огромни количества цифрови данни, стотици или хиляди входни набори могат да бъдат на разположение за решаване на проблем и създаване на модел. Чрез филтриране и подбор на най-подходящите входни променливи могат да се получат много ползи като по-добро разбиране на данните, по-лесна визуализация, намаляване на времето и ресурсите, необходими за обучение и използване на модела [3]. Не на последно място, входните променливи трябва да бъдат добри в описанието на данните и да носят достатъчно информация.

Подборът на данни за обработка ще се извърши спрямо предварително подготвени шаблони за входни данни, зависимост от източниците. По-долу е описан такъв шаблон на подбор на данни от LMS с отворен код Moodle.

Експерименталните данни са получени от система за управление на обучението (learning management system). Наборът от данни съдържа 63774 инстанции или образци, характеризиращи се със 7 атрибута, както следва: Време, контекст на събитието, Компонент, Име на събитието, Описание, Произход, IP адрес. Има някои случаи, които съдържат една липсваща (т.е. недостъпна) стойност на атрибут. За да се приложи алгоритъм за машинно обучение към набора от данни за обучение, е необходима предварителна обработка. Фазата на предварителна обработка на набор от данни се състои от два процеса: извличане на

функции и намаляване на набора от данни. Три атрибута от набора от данни са променени. Атрибутите Време и Описание не са необходими в процеса на моделиране и те се премахват. Освен това някои екземпляри се премахват от набора данни поради липсващи стойности.

Основните от дневниците за дейности са името на събитието и начина на класифицирането му според две основни групи потребители (ученик и учител). Ето защо ние избираме за гледна точка на учащия следните елементи:

- Разгледан модул на курса;
- Преглед на курса;
- Разгледана дискусия;
- Прегледан потребителски доклад за степен.

Според учителя важни елементи са:

- Елементът е създаден;
- Създаден е модул на курса;
- Актуализиран е модулът на курса;
- Преглед на отчета за дейността;
- Преглед на потребителски списък
- Потребителски профил гледан.

Изискването за клас трябва да бъде от типа на цяло число. Следователно, атрибутът на Името на събитието на всеки запис в набора от данни е променен, както в таблица 1.

Таблица 1. Име на събитието промяна на атрибут

No	Original Value	Formatted Value
1	Гледан модул на курса	0
2	Преглед на курса	1
3	Разгледана дискусия	2
4	Прегледан потребителски доклад за оценка	3
5	Елементът е създаден	4
6	Модулът на курса е създаден	5
7	Модулът на курса е актуализиран	6
8	Преглед на отчета за дейността	7
9	Преглед на потребителски списък	8
10	Потребителски профил е гледан	9

Примерният изглед на оригиналния неформатиран набор от данни е показан на фиг. 6, където всеки екземпляр е представен от ред, свързан със стойността на атрибутите.

"2/11/19, 11:15", Курс: Езици на програмиране, Система, Преглед на курса, Потребителят с id '7160' е видял курса с id '49'., Web, 193.57.20.13

"3/11/19, 11:32", Файл: Лекция, Файл, модул на курса, гледан, Потребителят с идентификатор „2“ е видял активността „ресурс“ с идентификатор на модула на курса '708'., Web, 212.5.158.162

"3/11/19, 11:32", Курс: Езици на програмиране, Доклад за дейността, Прегледан доклад за дейността, Потребителят с идентификатор „2“ е видял отчета за курса с id '49'., Web, 212.5.158.162

Фиг. 6. Пример за оригинален неформатиран набор от данни

Пример за форматирани набори от данни след предварителна обработка е показан на фиг.7

Курс: Езици на програмиране, система, 1, уеб
Файл: Лекция, Файл, 1, уеб
Курс: Езици на програмиране, доклад за дейността, 7, уеб

Фиг. 7. Пример за CSV форматирано въвеждане на данни след кодиране

4. Предварителна обработка на данни за обучение и валидиране на набор данни.

Аналитичният модел се създава след изпълнение на процеса на извличане на функции и намаляване на набора от данни и валидиран, използвайки набора от данни за валидиране. В резултат на това се създават различни модели за класификация и клъстеринг, които се използват за изграждане на работни процеси на analytics.

Фазата на обучение е насочена към създаване на хранилище на модели с използване на набор от данни за обучение и прилагане на класификационен алгоритъм за машинно обучение. Аналитичните модели се създават след изпълнение на процеса на извличане на функции и намаляване на набора от данни. Алгоритъмът за обучение за класификация на многокласност се основава на метода усреднен Perceptron [4]. Методът на усреднен Perceptron е проста версия на невронна мрежа. При този подход входовете се класифицират на няколко възможни изхода въз основа на линейна функция и след това се комбинират с набор от тегла, които се извличат от вектора на характеристиките. Perceprtons са бързи и поради това, че обработват случаи серийно, perceptrons могат да се използват с непрекъснато обучение [5].

Точността е една от критериите за оценка на производителността на модела и се изчислява като брой правилно класифицирани елементи, разделен на общия брой елементи в тестовия набор. Диапазонът на точност е от 0 (най-малко точен) до 1 (най-точен). Формулата за изчисляване на точността е:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \quad (1)$$

Измерените резултати за точност с избрания набор от характеристики са дадени в таблица 2I.

Таблица 2. Резултати от изпълнението

Number of Attributes	**Features set**	**Best Accuracy**
4	{Event context, Component, Origin, IP address }	88.01%

Използвана литература

[1] Simi M S, Mrs. Sankara Nayaki K, Dr.M.Sudheep Elayidom, "An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity", IOP Conf. Series: Materials Science and Engineering 225 (2017) 012279 doi:10.1088/1757-899X/225/1/012279

[2] Stuart Morton, Malika Mahoui, P. Joseph Gibson, Saidaiah Yechuri, "An Enhanced Utility-Driven Data Anonymization Method", TRANSACTIONS ON DATA PRIVACY, August 2012, 469 - 503

[3] MEHTA, Brijesh B.; RAO, Udai Pratap. Improved l-Diversity: Scalable Anonymization Approach for Privacy Preserving Big Data Publishing. *Journal of King Saud University-Computer and Information Sciences*, 2019.

[4] WANG, Li-E.; LI, Xianxian. A graph-based multifold model for anonymizing data with attributes of multiple types. *Computers & Security*, 2018, 72: 122-135.

[5] KAKATKAR, Chinmay; SPANN, Martin. Marketing analytics using anonymized and fragmented tracking data. *International Journal of Research in Marketing*, 2019, 36.1: 117-136.