

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

**РАЗРАБОТКА ПРИНЦИПОВ СОЗДАНИЯ ЭЛЕКТРОННОГО КУРСА
ПО ДИСЦИПЛИНЕ «BIGDATA В МЕДИЦИНЕ» И ИХ РЕАЛИЗАЦИЯ**

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 2 курса 220 группы

Направление подготовки 44.04.01 — Педагогическое образование

Факультета физико-математических и естественно-научных дисциплин

Тарасова Олега Игоревича

Научный руководитель

Д.т.н.

А.С. Фалькович

Зав. кафедрой:

К.ф.-м.н., доцент

М.В. Огнева

Саратов 2025

Введение

Актуальность темы. В современном мире многие аспекты человеческой жизни постепенно так или иначе движутся в сторону цифровизации. И медицина так же не стала исключением. Это прежде всего выражается в накоплении разного рода медицинских данных в виде электронных таблиц.

В связи с ежегодным увеличением темпов роста накапливаемых медицинских данных возникает потребность в их обработке и структуризации. Для этого существует научная дисциплина, называемая «BigData».

Что такое «BigData»? «Big Data» или большие данные — это структурированные или неструктурированные массивы данных большого объема. Их обрабатывают при помощи специальных автоматизированных инструментов, чтобы использовать для статистики, анализа, прогнозов и принятия решений.

Изучение дисциплины «BigData» обучающимися в медицинских вузах является одной из важнейших задач современного образовательного процесса в высших медицинских образовательных учреждениях, поскольку обучение данной дисциплине позволит учащимся работать с большими объёмами медицинских данных, более продуктивно заниматься обработкой различных данных.

Это в свою очередь позволит медицинским работникам более точно назначать лечение пациентам, ориентируясь на самые незначительные симптомы заболевания, ускорить процесс обследования каждого отдельно взятого пациента, а также полностью освободит медицинских работников от обследования пациентов, болеющих так называемыми «сезонными заболеваниями».

В связи с вышеперечисленными факторами, в медицине возникает необходимость в подготовке работников, специализирующихся на обработке больших данных.

Объект исследования: Big Data в медицине

Предмет исследования: преподавание больших данных студентам факультета фундаментальной медицины и медицинских технологий.

Гипотеза: изучение методов анализа больших данных осуществляется более успешно, если студенты первоначально решают задачи при помощи электронных таблиц, с дальнейшим переходом к готовым программным средствам, что позволяет лучше понять смысл совершаемых действий

Практическая значимость магистерской работы. Разработанный курс используется в преподавании дисциплины «Big Data в медицине» в Саратовском государственном университете на факультете фундаментальной медицины и медицинских технологий.

Цель: разработать принципы создания электронного курса по дисциплине «Big Data в медицине» и реализовать их на платформе Moodle.

Для достижения данной цели необходимо решить следующие задачи:

1. Провести анализ научной и методической литературы по теме исследования.
2. Определить оптимальные методы и технологии обучения работе с большими данными для студентов медицинских специальностей, а также выбрать наиболее подходящие инструменты для реализации электронного курса.
3. Разработать модель электронного курса, которая будет учитывать специфику дисциплины «Big Data в медицине» и соответствовать современным требованиям к качеству онлайн-образования.

4. Разработать и разместить их на платформе Moodle задания и контрольные вопросы для дисциплины «Big Data в медицине»

Выпускная квалификационная работа состоит из введения, пяти глав (Анализ литературы по теме исследования, Основополагающие процессы дисциплины «Big Data в медицине», Архитектура системы обработки Больших данных, Курс «Big Data в медицине» в Саратовском университете, Разработка заданий для дисциплины «Big Data в медицине»), и заключения.

Основное содержание работы

Основополагающими процессами связанными с Big Data в медицине являются:

1. Создание данных
2. Обслуживание данных
3. Синтез данных
4. Использование данных
5. Публикация данных
6. Архивация данных

Для работы с Большими данными используются сложные системы, в которых можно выделить несколько компонентов или слоёв. Обычно выделяют четыре уровня компонентов таких систем: прием, сбор, анализ данных и представление результатов. Это деление является в значительной мере условным так как, с одной стороны, каждый компонент в свою очередь может быть разделен на подкомпоненты, а с другой некоторые функции компонентов могут перераспределяться в зависимости от решаемой задачи и используемого программного обеспечения, например, выделяют хранение данных в отдельный слой.

Для работы с Большими данными разработчиками систем создаются модели данных, содержательно связанные с реальным миром. Разработка адекватных моделей данных представляет собой сложную аналитическую задачу, выполняемую системными архитекторами и аналитиками. Модель данных позволяет создать математическую модель взаимодействий объектов реального мира и включает в себя описание структуры данных, методы манипуляции данными и аспекты сохранения целостности данных.

Для хранения данных используются распределенные системы различных типов. Это могут быть файловые системы, базы данных, журналы, механизмы доступа к общей виртуальной памяти. Большинство систем хранения ориентированы исключительно на работу с Большими данными, они имеют крайне ограниченное число функций, (например, может отсутствовать возможность не только модификации, но и удаления поступивших данных) что объясняется внутренней сложностью создания высокоэффективных распределенных систем.

Для того, чтобы работа с данными происходила быстрее, системы хранения и обработки данных распараллеливаются в кластере (группа компьютеров, объединенных сетью для выполнения единой задачи). Однако, согласно гипотезе Брюера невозможно обеспечить одновременную согласованность (непротиворечивость) данных, доступность данных и устойчивость системы к отделению одиночных узлов. [9],[21]

Принципы электронного курса «Big Data в медицине»:

1. Принцип систематичности и доступности – реализуется за счет того, что последовательно рассматриваются процессы накопления, хранения, преобразования и анализа данных
2. Связь теории с практикой – реализуется за счет того, что в задачах рассматриваются реальные данные медицинской статистики
3. Принцип наглядности – реализуется за счет того, что задачи классификации и регрессии первоначально решаются при помощи электронных таблиц, с дальнейшим переходом к готовым программным средствам, что позволяет студентам лучше понять смысл совершаемых действий
4. Принцип решения новых задач – реализуется за счет того, что студенты активно вовлекаются в процесс решения задач анализа больших данных, применяя имеющиеся навыки работы в электронных таблицах и знания,

полученные на лекциях по данной дисциплине и дисциплинам медицинского цикла.

В качестве примера работы с большими данными студентам медицинского факультета были даны следующие задания:

Задание №1: Преобразование формата и разведочный анализ.

1. Открыть сайт <https://www.kaggle.com/>
2. Найти и загрузить датасет Brain stroke prediction dataset.

Для этого необходимо войти в Kaggle с аккаунтом Google. Скопировать также текстовое описание файла

3. Распаковать архив, просмотреть файл с расширением .csv.
4. **Закрыть** этот файл и **открыть с помощью** блокнота (Notepad).
5. Выполнить в блокноте следующие действия:

Заменить все запятые на точки с запятой;

Заменить все точки на запятые;

Сохранить файл **КАК** текстовый (с расширением .txt)

6. Закрыть файл, убедиться в том, что он существует в папке.
7. Открыть EXCEL, затем открыть текстовый **ИЗ** EXCEL (файл с разделителями, разделитель – точка с запятой). Сохранить **КАК** книгу EXCEL (). Закрыть файл, убедиться в том, что он существует в папке.
8. Открыть созданный файл с расширением .xlsx. Создать копию листа.
Работать только с копией листа!
9. Добавить в начало пустой столбец и пронумеровать строки
10. На копии листа создать фильтр: выделить все данные и на вкладке «данные» нажать кнопку «фильтр»

11. Пользуясь фильтром, проанализировать бинарные признаки (gender, hypertension, heart_disease, ever_married, Residence_type, stroke). Составить таблички распределения по полу всех остальных бинарных признаков и по инсульту всех остальных бинарных признаков.
12. Составить таблицу относительного распределения по инсульту всех остальных бинарных признаков.

Задание №2: Анализ категориальных и числовых признаков.

Категориальные признаки

1. С помощью фильтра составить таблицы абсолютного и относительного распределения каждого из категориальных признаков (work_type, smoking_status, residence_type) по инсульту:

общую

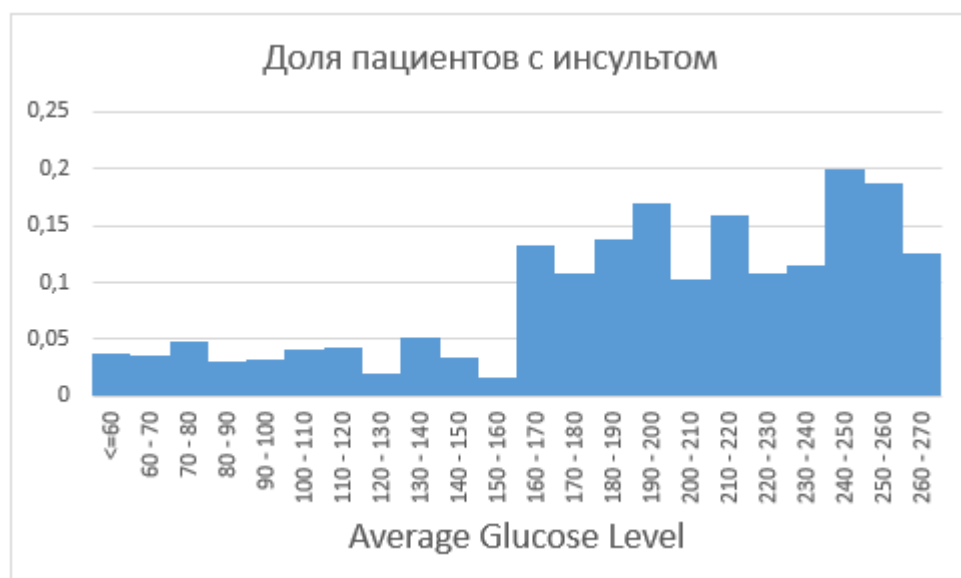
для мужчин

для женщин

Числовые признаки (avg_glucose_level и bmi)

2. Отсортировать данные по параметру «stroke» по убыванию
3. Найти минимальное и максимальное значения числовых признаков avg_glucose_level и bmi отдельно при stroke = 1 и stroke = 0.
4. С помощью функции =ЧАСТОТА() построить вариационные ряды и гистограммы параметров avg_glucose_level и bmi при stroke = 1 и stroke = 0.
5. Для этого:
 - а. напечатать столбец верхних границ интервалов для вариационного ряда: для avg_glucose_level рекомендуются от 60 до 270 с шагом 10 (всего 22 числа)

- б. В соседнем столбце выделить 23 ячейки, то есть на одну ячейку больше, чем границ интервалов, так как в последнюю ячейку запишется количество пациентов с показателем $AGL > 270$
 - в. Вызвать функцию `=ЧАСТОТА( )`. В качестве массива данных выбрать нужные ячейки с значениями содержания глюкозы у пациентов с инсультом, в качестве массива интервалов – сформированный массив верхних границ интервалов. после заполнения окон функции «ЧАСТОТА» нажать сочетание клавиш CTRL SHIFT ENTER (а не OK или ENTER)
6. В следующем столбце построить вариационный ряд для параметров `avg_glucose_level` у пациентов без инсульта.
7. Построить гистограммы для вариационных рядов параметра `avg_glucose_level` при `stroke = 1` и `stroke = 0`.
8. В следующем столбце вычислить доли пациентов с инсультом в каждом интервале вариационного ряда: в каждой строке количество пациентов с инсультом с данным уровнем глюкозы поделить на суммарное количество пациентов с этим уровнем глюкозы (при `stroke = 1` и `stroke = 0`). Построить гистограмму по вычисленным данным. Должна получиться такая гистограмма.



9. Подписи горизонтальной оси можно сформировать в следующем столбце с помощью функции =СЦЕПИТЬ()
10. Найти в интернете верхнюю границу нормы содержания сахара в крови в мг/100 мл и сравнить с гистограммой.
11. Построить аналогичным способом вариационные ряды и гистограммы для индекса массы тела и возраста.

Задание №3: Кодирование и нормализация признаков.

1. После столбца “gender” вставить 2 столбца и назвать их “male” и “female”. Заполнить эти столбцы единицами и нулями.
2. После столбца “Residence_type” вставить 2 столбца и назвать их “Urban” и “Rural”. Заполнить эти столбцы единицами и нулями.
3. Перед столбцом “stroke” вставить 4 столбца и назвать их “WT_bin”, “AGL_bin”, “BMI_bin” и “SMOKE_bin”.
4. Столбец “WT_bin” заполнить:
 - 0, если Children
 - 3, если Self-employed
 - для мужчин: 1, если Govt_job, и 2, если Private
 - для женщин: 1, если Private, и 2, если Govt_job
5. Столбец “SMOKE_bin” заполнить:
 - 1, если Unknown
 - 0, если never smoked
 - 1, если smokes или formerly smoked
6. Столбец “AGL_bin” заполнить 0, если avg_glucose_level меньше или равно 160, и 1, если avg_glucose_level больше 160

7. Столбец “BMI_bin” заполнить 0, если bmi меньше или равно 27, и 1, если bmi больше 27
8. Столбец “AGE_bin” заполнить 0, если age меньше или равно 55, и 1, если age больше 55
9. Подсчитать сумму кодированных признаков отдельно для stroke = 1 и для stroke = 0. Подсчитать частоты и построить гистограммы.

Задание № 4: Дерево решений

1. С помощью инструмента «Фильтр» подсчитать неопределенность Джини (Gini impurity)

$$\Pr(P=0) \cdot \Pr(Y=0|P=0) \cdot \Pr(Y=1|P=0) + \Pr(P=1) \cdot \Pr(Y=0|P=1) \cdot \Pr(Y=1|P=1)$$
 для показателей age, hypertension, heart_disease, avg_glucose_level, bmi, smoking_status и для кодированного показателя WT_bin

Пример расчета неопределенности Джини для показателя age

Количество пациентов с данными показателями (абсолютные числа)

age		Stroke (1)	No stroke (0)
<=55 (0)	3304	38	3266
>55 (1)	1677	210	1467
все возраста	4981	248	4733

Вероятности $\Pr(P=0)$ и $\Pr(P=1)$: $P=0$ – возраст ≤ 55 ; $P=1$ – возраст > 55 ;

$\Pr(P=0) \ 3304/4981 = 0,663321$

$\Pr(P=1) \ 1677/4981 = 0,336679$

Условные вероятности $\Pr(Y=0|P=0)$ и $\Pr(Y=1|P=0)$: $Y=0$ – нет инсульта, $Y=1$ – инсульт при условии, что возраст ≤ 55

$$\Pr(Y=0|P=0) 3266/3304 = 0,988499$$

$$\Pr(Y=1|P=0) 38/3304 = 0,011501$$

Условные вероятности $\Pr(Y=0|P=1)$ и $\Pr(Y=1|P=1)$: $Y=0$ – нет инсульта, $Y=1$ – инсульт при условии, что возраст >55

$$\Pr(Y=0|P=1) 1467/1677 = 0,874776$$

$$\Pr(Y=1|P=1) 210/1677 = 0,125224$$

Произведения

$$\Pr(P=0)*\Pr(Y=0|P=0)*\Pr(Y=1|P=0) = 0,663321 * 0,011501 * 0,988499 = 0,007541$$

$$\Pr(P=1)*\Pr(Y=0|P=1)*\Pr(Y=1|P=1) = 0,336679 * 0,125224 * 0,874776 = 0,036881$$

$$\text{Неопределенность Джини } 0,007541 + 0,036881 = 0,044422$$

		Y=1	Y=0	Произведения
$\Pr(P=0)$	0,663321	0,011501	0,988499	0,007541
$\Pr(P=1)$	0,336679	0,125224	0,874776	0,036881

2. Выбрать показатель с наименьшей неопределенностью Джини – это будет первое условие ветвления.
3. Подсчитать, сколько пациентов с инсультом и без инсульта попадет в каждую ветвь

Заключение

В результате выполнения работы проведен анализ научной и методической литературы по теме исследования. Анализ рассмотренных источников показал, что методика преподавания Больших Данных находится в стадии становления. Наметилось две основных тенденции в преподавании этой дисциплины: хранение больших данных (системы управления базами данных) и анализ данных на основе искусственного интеллекта. Выявлены основополагающие процессы дисциплины «Big Data в медицине», разработаны задания для электронного курса, учитывающие специфику дисциплины «Big Data в медицине» и соответствующие современным требованиям к качеству онлайн-образования, определены оптимальные методы и технологии обучения работе с большими данными для студентов медицинских специальностей. Определены основные принципы и направления создания электронного курса по дисциплине «Big Data в медицине».

На основе изученных материалов был создан курс, нацеленный на изучение BigData. Данный курс ориентирован в основном на студентов высших медицинских образовательных учреждений, но подойдет и для обучающихся, не связанных с медициной.

Отдельные части магистерской работы были представлены на конференции «Информационные технологии в образовании-2024» и опубликованы:

Тарасов О.И. Разработка методических принципов создания электронного курса по дисциплине «Big Data в медицине» / О.И. Тарасов, А.С. Фалькович // Информационные технологии в образовании: сборник научных статей – Саратов: Саратовский университет, 2024. – Вып. 7. С. 292-295