МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А.И. ГЕРЦЕНА»

Направление подготовки

09.03.01 – Информатика и вычислительная техника

Профиль «Технологии разработки программного обеспечения»

ИНДИВИДУАЛЬНАЯ ЗАДАЧА

по дисциплине «Анализ данных и основы Data Science»

Работу выполнил студент 2 курса 2-1 группы:

Стецук Максим

САНКТ-ПЕТЕРБУРГ

2023

Условие задачи

Был проведён анонимный опрос среди студентов двух групп (по 20 человек в каждой), обучающихся в различных сферах подготовки. Первая группа студентов обучается в колледже в сфере информационных технологий, а вторая в колледже в медицинской сфере. В ходе опроса каждому студенту предоставлялся список из 10 наиболее развивающихся и известных ИИ, предназначенных для различного рода задач. Им необходимо было расположить представленные ИИ в порядке от наиболее полезного/известного для конкретного студента, до наименее известного (соответственно: значение 1 - очень полезный/известный для конкретного студента, например он использует данный ИИ с некоторой периодичностью, значение 10 - наименее известный, имеет минимальный опыт использования, никогда не использовал или даже не слышал о нём). Результаты опросов для двух групп представлены в таблицах:

IT специальность

№	ChatGPT	CodeGPT	BLOOM	Whisper	DALL-E 2	Craiyon	Stable Diffusion	Imagen	Make-A-Video	GitHub Copilot
1	1	3	4	8	7	5	6	10	9	2
2	1	2	10	5	4	6	7	8	9	3
3	3	2	9	8	7	4	10	5	6	1
4	2	3	9	7	10	8	6	1	4	5
5	2	1	4	10	5	6	9	7	8	3
6	1	3	6	5	9	8	4	10	7	2
7	2	4	8	7	1	6	3	9	10	5
8	3	4	6	8	10	9	7	1	2	5
9	4	3	8	5	9	1	2	10	7	6
10	1	4	3	2	9	7	8	6	10	5
11	5	10	7	6	1	4	9	2	3	8
12	2	5	9	8	3	10	7	1	4	6
13	1	2	5	7	6	4	3	8	10	9
14	1	3	4	10	2	9	8	5	7	6
15	1	4	7	6	8	3	10	9	5	2
16	1	5	10	9	2	8	4	7	6	3
17	2	3	7	8	6	10	9	5	1	4
18	2	3	8	9	1	4	5	10	7	6
19	1	6	8	2	3	9	4	10	7	5
20	1	7	2	5	10	4	8	6	9	3

Медицинская специальность

№	ChatGPT	CodeGPT	BLOOM	Whisper	DALL-E 2	Craiyon	Stable Diffusion	Imagen	Make-A-Video	GitHub Copilot
1	9	6	3	5	10	4	7	1	2	8
2	3	4	9	7	2	10	6	5	8	1
3	7	10	8	4	1	6	5	9	2	3
4	5	4	8	9	6	7	1	3	10	2
5	5	10	3	9	1	7	4	6	8	2
6	9	8	10	2	7	1	4	6	3	5
7	3	2	5	9	1	6	7	8	4	10
8	3	1	2	8	7	10	5	6	9	4
9	1	6	2	8	7	10	5	3	9	4
10	1	6	10	4	7	8	2	5	3	9
11	4	7	1	3	5	10	2	9	6	8
12	1	7	3	2	9	10	5	6	8	4
13	5	2	6	1	10	9	3	4	8	7
14	10	1	4	5	6	8	7	3	2	9
15	3	6	2	10	8	7	1	5	9	4
16	1	3	8	10	5	9	2	4	6	7
17	8	3	2	4	6	1	9	5	7	10
18	7	2	1	10	4	9	3	8	5	6
19	2	9	5	4	1	8	7	3	6	10
20	4	6	10	2	7	8	9	1	5	3

С помощью корреляционного анализа необходимо сравнить совокупности ответов студентов каждой из групп, а также построить и изобразить графически (полигон, кумулянта и эмпирическая функция распределения) дискретный вариационный ряд для каждой группы студентов по выборке оценок для 3 ИИ: ChatGPT, CodeGPT, GitHub Copilot. Проанализировать полученный результат, сделать выводы о востребованности прикладных ИИ в двух различных сферах подготовки студентов.

Решение задачи

Корреляционный анализ для совокупностей ответов студентов двух групп

Для проведения корреляционного анализа и построения корреляционного поля проведём ранжирование средних результатов для каждой группы.

В каждой таблице найдём среднее значение оценки для каждого ИИ (для каждого столбца), а затем проведём ранжирование от меньшего к большему (т.к. по условию 1 - соответствует наиболее полезному/известному ИИ для конкретного студента, а 10 - наименее полезному/известному для конкретного студента, то при ранжировании средних значение, наименьшее значение будет иметь ранг 1, а наибольшее - 10).

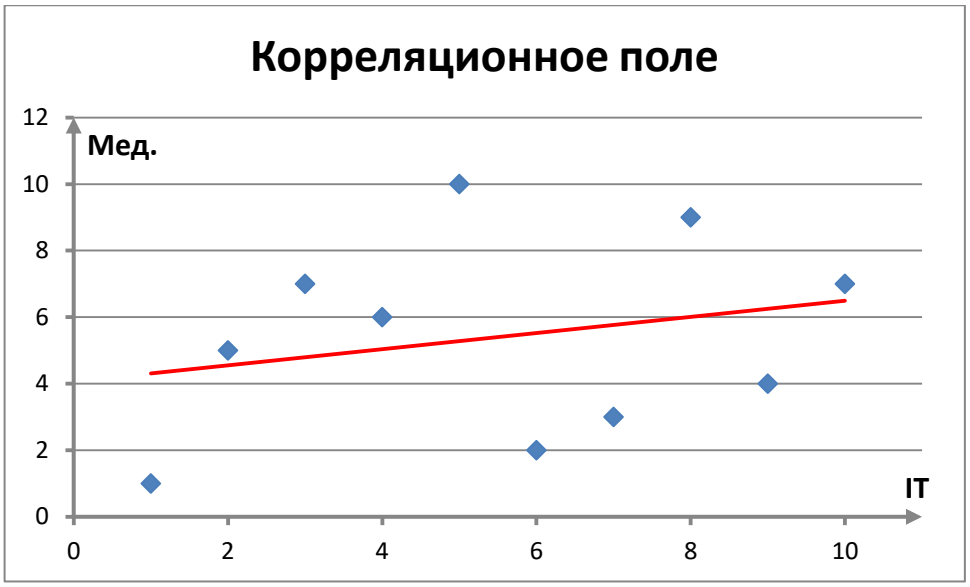
Таб. IT	ChatGPT	CodeGPT	BLOOM	Whisper	DALL-E 2	Craiyon	Stable Diffusion	Imagen	Make-A-Video	GitHub Copilot
1	1	3	4	8	7	5	6	10	9	2
2	1	2	10	5	4	6	7	8	9	3
3	3	2	9	8	7	4	10	5	6	1
4	2	3	9	7	10	8	6	1	4	5
5	2	1	4	10	5	6	9	7	8	3
6	1	3	6	5	9	8	4	10	7	2
7	2	4	8	7	1	6	3	9	10	5
8	3	4	6	8	10	9	7	1	2	5
9	4	3	8	5	9	1	2	10	7	6
10	1	4	3	2	9	7	8	6	10	5
11	5	10	7	6	1	4	9	2	3	8
12	2	5	9	8	3	10	7	1	4	6
13	1	2	5	7	6	4	3	8	10	9
14	1	3	4	10	2	9	8	5	7	6
15	1	4	7	6	8	3	10	9	5	2
16	1	5	10	9	2	8	4	7	6	3
17	2	3	7	8	6	10	9	5	1	4
18	2	3	8	9	1	4	5	10	7	6
19	1	6	8	2	3	9	4	10	7	5
20	1	7	2	5	10	4	8	6	9	3
Сумма	37	77	134	135	113	125	129	130	131	89
Среднее	1,85	3,85	6,7	6,75	5,65	6,25	6,45	6,5	6,55	4,45
РАНГ	1	2	9	10	4	5	6	7	8	3

Таб. Мед	ChatGPT	CodeGPT	BLOOM	Whisper	DALL-E 2	Craiyon	Stable Diffusion	Imagen	Make-A-Video	GitHub Copilot
1	9	6	3	5	10	4	7	1	2	8
2	3	4	9	7	2	10	6	5	8	1
3	7	10	8	4	1	6	5	9	2	3
4	5	4	8	9	6	7	1	3	10	2
5	5	10	3	9	1	7	4	6	8	2
6	9	8	10	2	7	1	4	6	3	5
7	3	2	5	9	1	6	7	8	4	10
8	3	1	2	8	7	10	5	6	9	4
9	1	6	2	8	7	10	5	3	9	4
10	1	6	10	4	7	8	2	5	3	9
11	4	7	1	3	5	10	2	9	6	8
12	1	7	3	2	9	10	5	6	8	4
13	5	2	6	1	10	9	3	4	8	7
14	10	1	4	5	6	8	7	3	2	9
15	3	6	2	10	8	7	1	5	9	4
16	1	3	8	10	5	9	2	4	6	7
17	8	3	2	4	6	1	9	5	7	10
18	7	2	1	10	4	9	3	8	5	6
19	2	9	5	4	1	8	7	3	6	10
20	4	6	10	2	7	8	9	1	5	3
Сумма	91	103	102	116	110	148	94	100	120	116
Среднее	4,55	5,15	5,1	5,8	5,5	7,4	4,7	5	6	5,8
РАНГ	1	5	4	7	6	10	2	3	9	7

Получили ранжирование ответов для каждой группы. Вынесем полученные результаты в отдельную таблицу.

ИИ	ChatGPT	CodeGPT	BLOOM	Whisper	DALL-E 2	Craiyon	Stable Diffusion	Imagen	Make-A-Video	GitHub Copilot
Ранг IT	1	2	9	10	4	5	6	7	8	3
Ранг Мед	1	5	4	7	6	10	2	3	9	7

Построим корреляционное поле для рангов оценок двух групп.



Из графического представления зависимости можно сделать предположение, что взаимосвязь между оценками двух групп студентов является линейной с положительным направлением.

Определим разности рангов, их квадраты и сумму квадратов, затем найдём коэффициент ранговой корреляции Спирмена по формуле:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ИИ	ChatGPT	CodeGPT	BLOOM	Whisper	DALL-E 2	Craiyon	Stable Diffusion	Imagen	Make-A-Video	GitHub Copilot
Ранг IT	1	2	9	10	4	5	6	7	8	3
Ранг Мед	1	5	4	7	6	10	2	3	9	7
di = Xi - Yi	0	-3	5	3	-2	-5	4	4	-1	-4
di^2	0	9	25	9	4	25	16	16	1	16

n	10
sum di^2	121
rs	0,267

Значит у нас слабая теснота корреляционной связи для совокупностей ответов двух групп студентов (т.к. $0 < r < 0{,}299$).

Проверим, существует ли положительная корреляционная связь между оценками двух групп. Для этого используем t - статистику Стьюдента с $\nu = (n-2)$ степенями свободы:

$$t = |r_s|\sqrt{\frac{n-2}{1-r_s^2}} = |0{,}267|\sqrt{\frac{10-2}{1-0{,}267^2}} = 0{,}783$$

При уровне значимости α = 0.05 для односторонней (правосторонней) критической области:

$$t_{\text{кр}} = t_{0,05,8} = 1{,}86$$

$$t_{\text{расч}} = 0{,}783$$

Получается, $0{,}783 < 1{,}86\ (t_{\text{расч}} < t_{\text{кр}})$. Следовательно, связь между совокупностями результатов двух групп не является статистически значимой при 5 %-ном уровне значимости.

Поэтому можно сделать вывод, что мнения студентов двух групп (в совокупности) про 10 представленных в опросе искусственных интеллектов сильно различаются, и данное различие может быть вызвано непосредственно направлением их подготовки и различием потребности в определённых инструментах у каждого студента.

Анализ результатов опроса групп студентов по выборке из 3 ИИ

Рассмотрим ранги, которые выставили студенты 2 различных направлений подготовки для ChatGPT, CodeGPT и GitHub Copilot.

Таб. IT	ChatGPT	CodeGPT	GitHub Copilot
1	1	3	2
2	1	2	3
3	3	2	1
4	2	3	5
5	2	1	3
6	1	3	2
7	2	4	5
8	3	4	5
9	4	3	6
10	1	4	5
11	5	10	8
12	2	5	6
13	1	2	9
14	1	3	6
15	1	4	2
16	1	5	3
17	2	3	4
18	2	3	6
19	1	6	5
20	1	7	3

Таб. Мед	ChatGPT	CodeGPT	GitHub Copilot
1	9	6	8
2	3	4	1
3	7	10	3
4	5	4	2
5	5	10	2
6	9	8	5
7	3	2	10
8	3	1	4
9	1	6	4
10	1	6	9
11	4	7	8
12	1	7	4
13	5	2	7
14	10	1	9
15	3	6	4
16	1	3	7
17	8	3	10
18	7	2	6
19	2	9	10
20	4	6	3

Построим дискретный вариационный ряд по данным рангам для каждой группы студентов.

IT специальность

Xi	1	2	3	4	5	6	7	8	9	10
mi	12	12	13	6	8	5	1	1	1	1

Медицинская специальность

Xi	1	2	3	4	5	6	7	8	9	10
mi	7	6	8	8	4	6	6	4	5	6

Теперь построим полигоны для данных распределений.

Найдём накопленные частоты и частности для двух распределений.

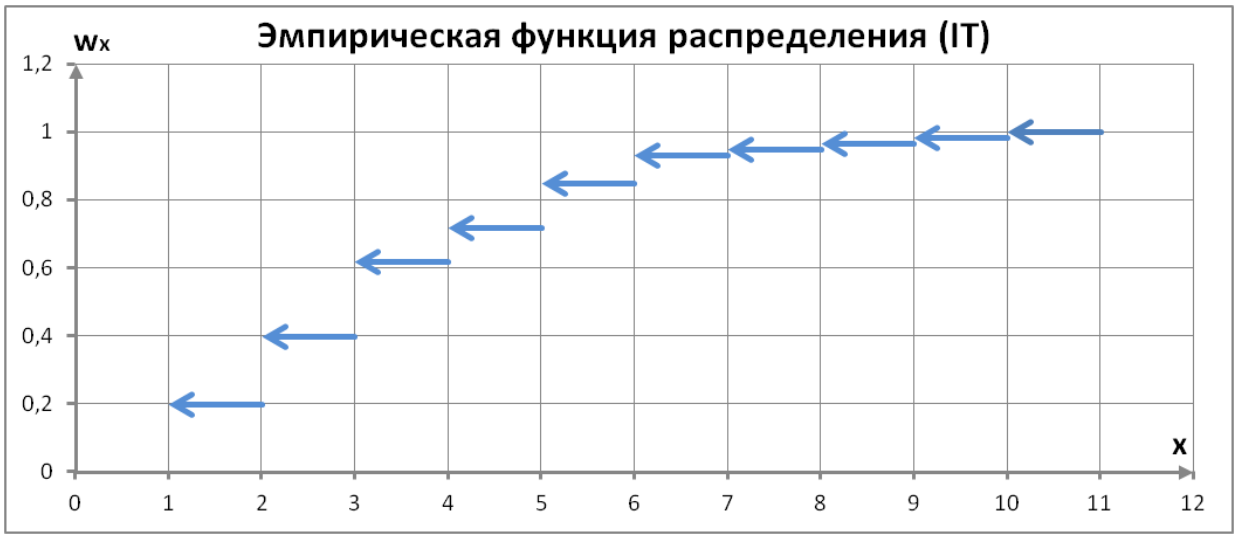
IT специальность

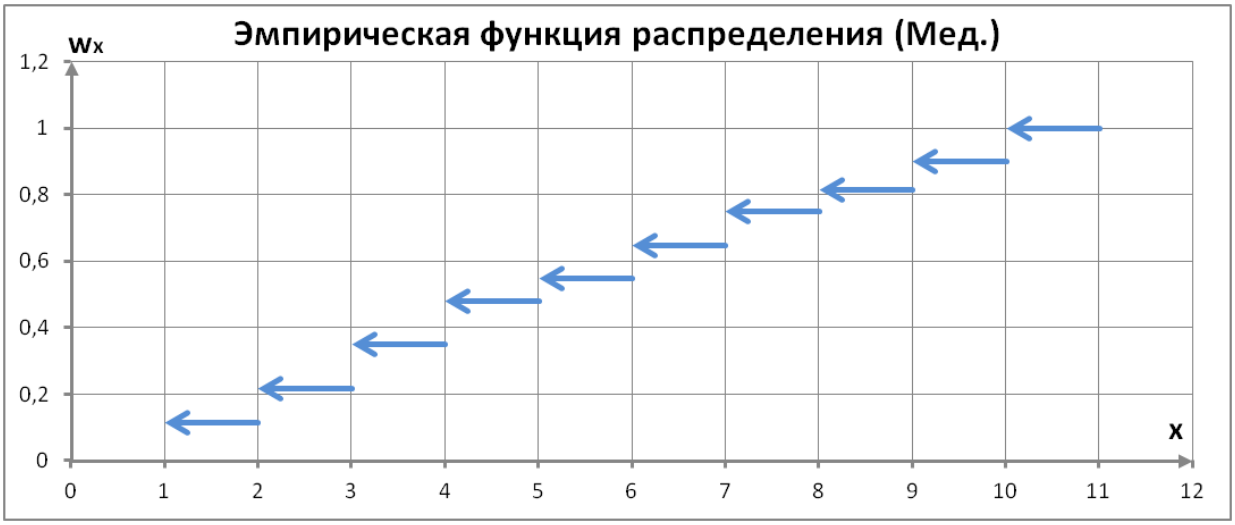
Xi	1	2	3	4	5	6	7	8	9	10	11
mxi	0	12	24	37	43	51	56	57	58	59	60
Wxi	0,000	0,200	0,400	0,617	0,717	0,850	0,933	0,950	0,967	0,983	1,000

Медицинская специальность

Xi	1	2	3	4	5	6	7	8	9	10	11
mxi	0	7	13	21	29	33	39	45	49	54	60
Wxi	0	0,117	0,217	0,350	0,48	0,550	0,650	0,75	0,817	0,900	1

По данным накоплениям построим кумулянты и эмпирические функции распределения для обеих специальностей.





Исходя из полученных дискретных вариационных рядов, полигонов распределений, кумулянт и эмпирических функций распределения видно, что у студентов, обучающихся по направлению информационных технологий, ChatGPT, CodeGPT и GitHub Copilot занимают лидирующую позицию среди остальных ИИ по известности/полезности, т.к. большая часть участников опроса указали оценку (ранг) 1, 2 или 3, ведь именно эти ИИ имеют большие возможности и спектр применений при программировании и т.п. В то же время у студентов, обучающихся на направлении связанном с медициной, оценки (ранги) сильно варьируются от 1 до 10.

Вывод: В данной работе мной были проанализированы результаты опроса двух групп студентов, обучающихся по различным направлениям подготовки. С помощью корреляционного анализа выяснено, что полученные результаты сильно отличаются друг от друга и связь между совокупностями результатов для двух групп не является статистически значимой. Причиной данного отличия можно считать более глубокую осведомлённость студентов обучающихся по направлению связанному с информационными технологиями об ИИ и поверхностные знания о них у студентов медицинского направления. Рассмотрев дискретные вариационные ряды, полигоны распределения, кумулянты и эмпирические функции распределения было выяснено, что прикладные ИИ (используемые в определённой области) более популярны у студентов с IT направления. Это объясняется тем, что такие ИИ, как ChatGPT, CodeGPT и GitHub Copilot существенно помогают при написании кода или разработке алгоритма, однако не способны существенно помочь студентам в изучении медицины, т.к. данное направление больше ориентировано на практическое применение (применение в реальном мире) полученных знаний.