ФГАОУ ВО «Волгоградский государственный университет»
Институт математики и информационных технологий
Кафедра информационных систем и компьютерного моделирования

Бут Александр Андреевич

«Использование классификаторов в машинном обучении»

Направление подготовки:	09.03.04 «Программная инженерия»
Группа:	ПРИб-201
Ответственный за организацию практики:	Корнаухова М.А., к.ф.-м.н., доцент каф. ИСКМ

Волгоград - 2022

Цель и задачи

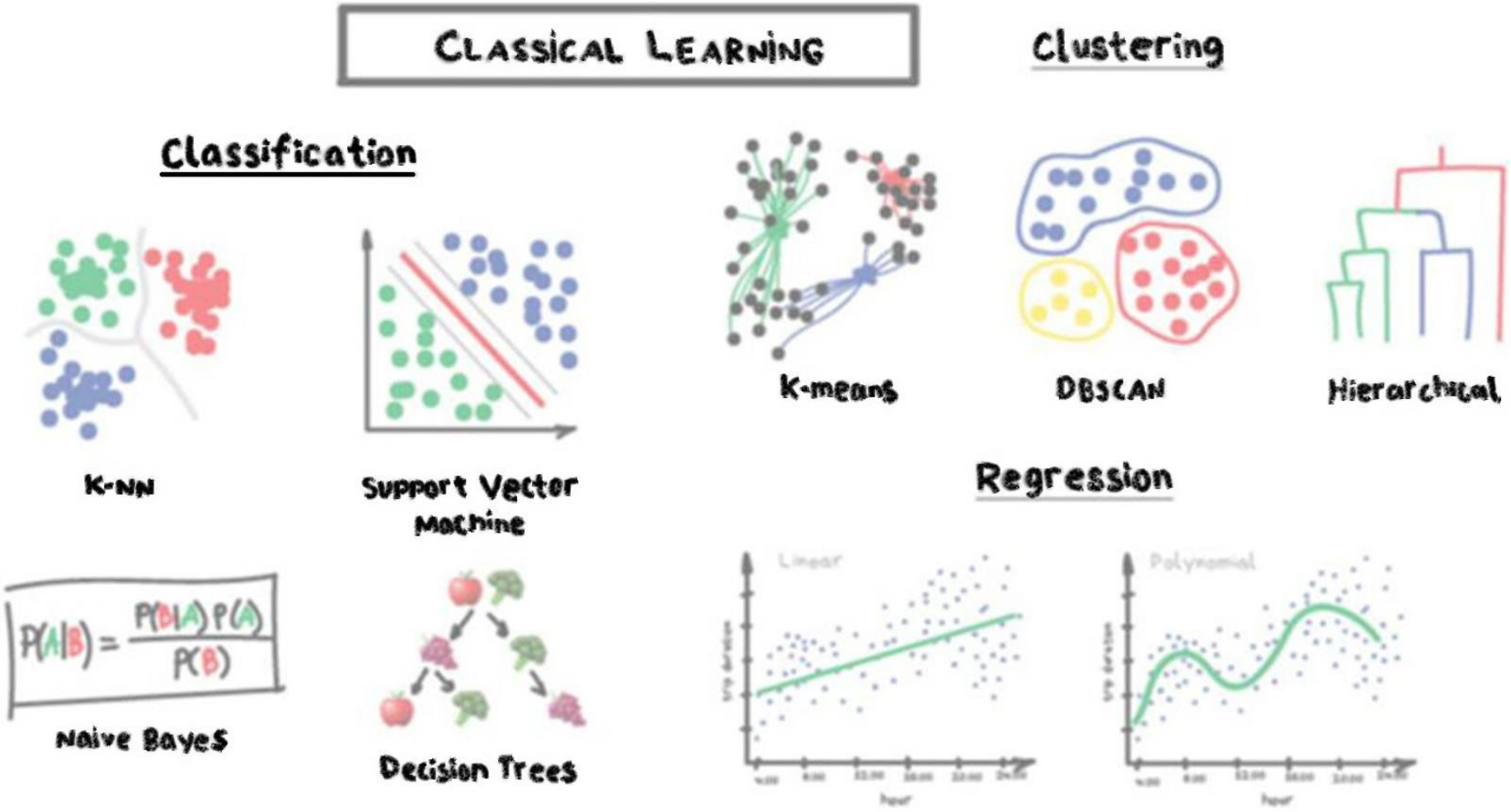
Цель:

Сравнительный анализ результатов работы нескольких классификаторов для решения задач анализа данных.

Основные задачи:

1. Сбор и изучение литературы по теме исследования.
2. Рассмотрение линейных моделей для классификации данных.
3. Программно реализовать каждую модель классификации.
4. Провести сравнительный анализ результатов работы нескольких классификаторов для решения задач анализа данных.

Основные алгоритмы машинного обучения



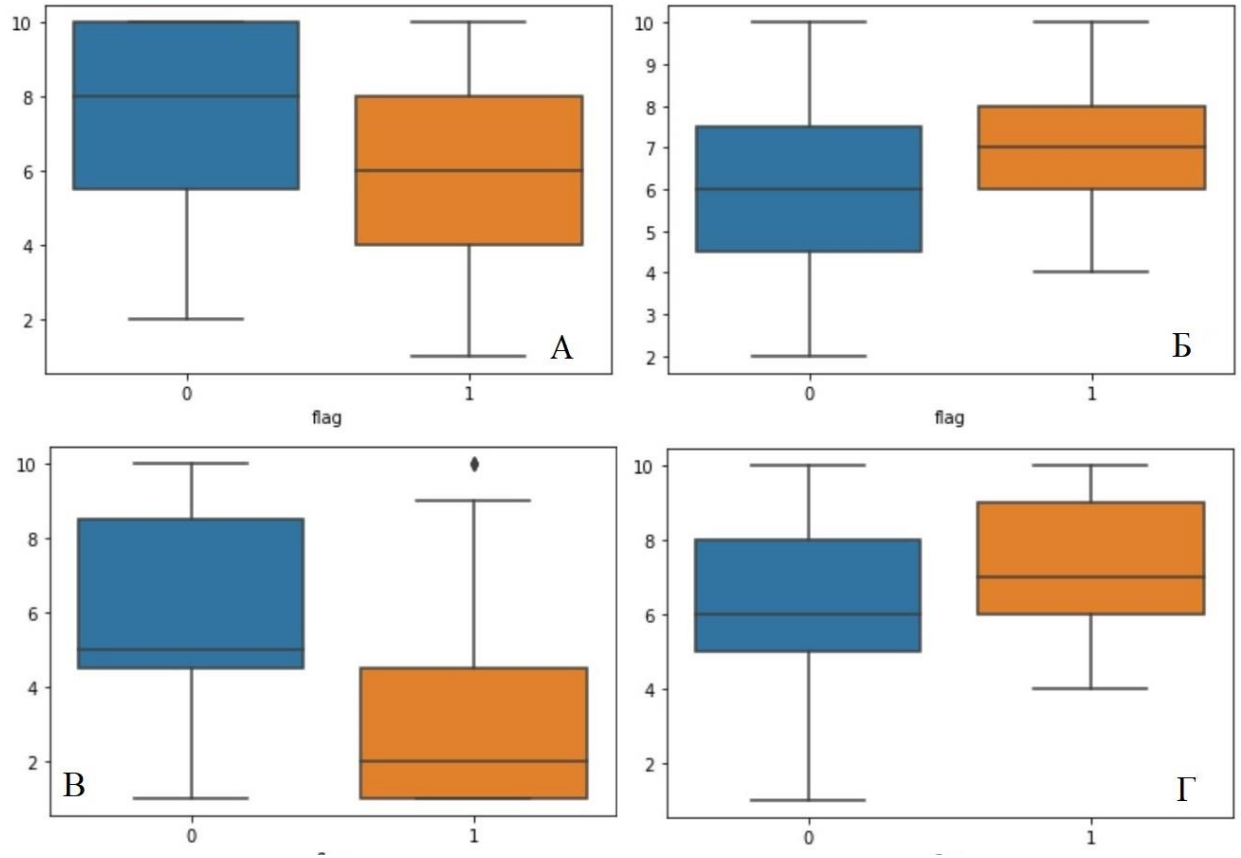
Шаги машинного обучения

Предобработка входных данных

1. ФИО	2. Ваше н	3. Сумм	4. Час	5. Ча	6. Нрави	7. Нрави	8. Нр	9. Ум	10. Нр	11. Нра	12. Н	13. Н	flag
Устарханс	МОС	213	10	3	6	5	5	10	1	7	1	8	1
Зенченко	МОС	168	8	5	6	5	5	3	4	8	1	8	1
Васильев	ПМФ	164	8	5	6	5	8	4	1	6	2	9	0
Гувалов Р	ИСТ	190	3	2	5	6	8	1	3	5	5	9	1
Устиновск	ПРИ	264	10	10	10	6	1	2	2	7	1	5	1
Горбачен	МОС	181	7	6	10	6	9	3	1	10	4	7	1
Панафиди	ИСТ	219	5	2	8	7	9	7	8	10	4	7	1
Омельчен	ПРИ	225	6	8	6	7	7	8	2	7	4	10	1
Олейник А	МОС	204	8	1	5	7	6	9	3	10	1	1	1
Зверяев Е	МОС	246	8	5	5	7	7	3	2	7	1	6	1
Пономаре	МОС	240	10	7	9	7	6	2	7	10	4	4	1
Савельев	МОС	196	8	5	6	7	9	2	1	9	3	8	1
Мыльнико	ПМФ	211	8	4	8	7	7	3	10	7	5	9	0
Лиджиев Е	ИСТ	220	9	4	7	8	10	9	6	9	6	7	1
Фролова Н	ИСТ	265	10	6	8	8	5	3	4	9	7	7	1
Агапченко	ИСТ	199	10	3	10	8	5	2	8	8	4	7	1
Резанов К	ПРИ	238	10	4	9	8	5	1	6	10	6	6	1
Астахов Д	ПРИ	210	10	1	9	8	5	6	8	10	5	8	1
Ракчеев Н	ПРИ	246	8	6	4	8	6	3	4	4	3	7	1
Чернышов	МОС	222	7	5	7	8	4	1	4	8	2	7	1
Крайнев Н	МОС	224	10	7	8	8	5	1	4	8	3	6	1
Макарова	МОС	222	3	10	5	8	5	3	1	2	1	5	1

Статистическая обработка данных

Построение диаграмм размаха



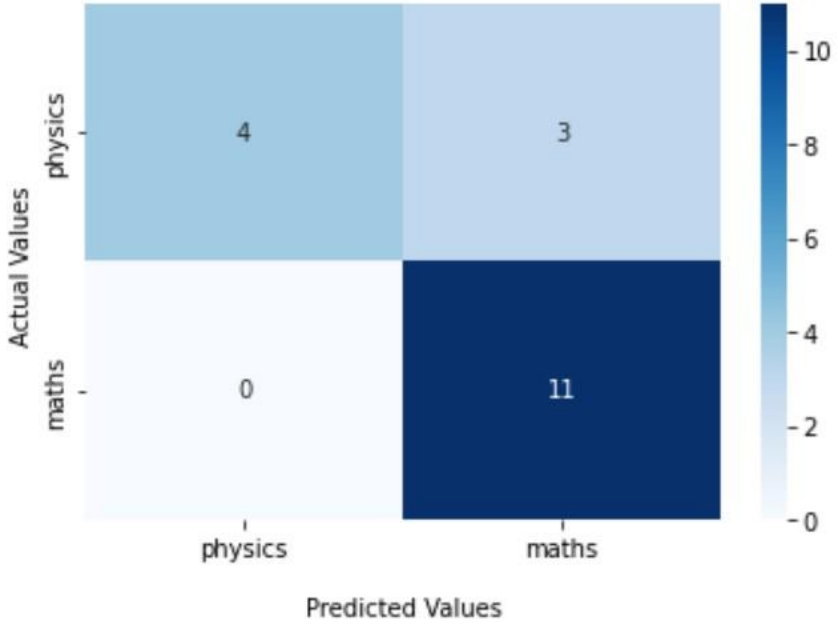
0 – физическое направление.
1 – математическое направление

А: «Нравится ли вам разбираться в принципе работы электронных устройств?»
Б: «Нравится ли вам решать математические задачи?»
В: «Нравится ли вам паять схемы?»
Г: Нравится ли вам решать головоломки?»

Матрица корреляций и формулы для расчёта статистических показателей

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

General number of each criteria



$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive + false negative}}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative + false positive}}$$

$$\text{Efficiency} = \sqrt{\text{Sensitivity} * \text{Sensitivity}}$$

Сравнительный анализ результатов для базы с исключенными элементами

	KNN	DT	RF	NB	LOGREG	SV
чувствительность	0,2646	0,4723	0,3262	0,5176	0,5314	0,4003
специфичность	0,7984	0,6298	0,7798	0,6761	0,7672	0,7780
эффективность	0,4136	0,5230	0,4739	0,5770	0,6257	0,4977

Наилучшая чувствительность = 0,5314;
Наилучшая специфичность = 0,7672;
Наилучшая эффективность = 0,6252;
Для метода LOGREG (логистическая регрессия)

Сравнительный анализ результатов для полной базы

	KNN	DT	RF	NB	LOGREG	SV
чувствительность	0,2998	0,6268	0,3185	0,7142	0,5428	0,5664
специфичность	0,7107	0,8737	0,7808	0,7272	0,7763	0,7146
эффективность	0,4269	0,7322	0,4647	0,7207	0,6295	0,6210

Наилучшая чувствительность = 0,6268;
Наилучшая специфичность = 0,8737;
Наилучшая эффективность = 0,7322;
Для метода DT (Деревья решений)

Заключение

В ходе работы были решены следующие задачи:
1. Была собрана и изучена литература по теме исследования.

2. Было рассмотрено несколько линейных моделей для классификации данных.

3. Каждая модель была программно реализована.

4. Был проведен сравнительный анализ результатов работы нескольких классификаторов для решения задач анализа данных.

Для метода деревьев решений была
достигнута эффективность 73%, для алгоритма наивного байеса 72%.