МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО «СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ
Директор НОЦ «Институт
непрерывного образования»
_____ Е.В. Мошкина
«12» февраля 2024 г.

ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА
ПРОФЕССИОНАЛЬНОЙ ПЕРЕПОДГОТОВКИ

«Программист - исследователь больших данных»

Форма обучения – очно-заочная.
Объем программы – 342 часа.

Красноярск 2024

УЧЕБНЫЙ ПЛАН
дополнительной профессиональной программы профессиональной переподготовки
«Программист - исследователь больших данных»

Форма обучения – очно-заочная (с использованием электронного обучения и дистанционных образовательных технологий).
Срок обучения – 342 часа.

№ п/п	Наименование модулей (дисциплин)	Общая трудоем-кость, часов	Всего контактных, часов	Контактные часы			СРС, часов	Формы Контроля
				Лекции	Лабора-торные работы	Практические и семинарские занятия		
1.	Методы анализа данных	108	54	18	-	36	54	Зачет
2.	Теория вычислительного обучения	144	72	18	-	54	72	Экзамен
3.	Инженерия машинного обучения	54	26	8	-	18	28	Зачет
4.	Итоговая аттестация	36	6	-	-	6	30	Защита итоговой аттестационной работы (проекта)
	Итого	342	158	44	-	114	184	

УЧЕБНО-ТЕМАТИЧЕСКИЙ ПЛАН
дополнительной профессиональной программы профессиональной переподготовки
«Программист - исследователь больших данных»

Категория слушателей: ***лица, имеющие/получающие высшее образование***
Срок обучения: 12 недель
Форма обучения: очно-заочная
Режим занятий: 2-3 часа в день

№ п/п	Наименование модулей (дисциплин)	Общая трудоем-кость, часов	Всего контактных часов	Контактные часы			СРС, часов	Результаты обучения
				Лекции	Лабора-торные работы	Практ. и семинарские занятия		
1	**Методы анализа данных**	**108**	**54**	**18**	**-**	**36**	**54**	**РО1-РО5**
1.1	Тема 1.1. Введение в анализ данных	4	2	2	-	-	2	РО1
1.2	Тема 1.2. Первичная обработка данных	14	6	2	-	4	8	РО2
1.3	Тема 1.3. Задача классификации	22	12	4	-	8	10	РО3-РО5
1.4	Тема 1.4. Задача восстановления регрессии	22	12	4	-	8	10	РО3-РО5
1.5	Тема 1.5. Деревья решений	16	8	2	-	6	8	РО3-РО5
1.6	Тема 1.6. Ансамбли моделей	16	8	2	-	6	8	РО3-РО5
1.7	Тема 1.7. Кластеризация	14	6	2	-	4	8	РО3-РО5
2	**Теория вычислительного обучения**	**144**	**72**	**18**	**-**	**54**	**72**	**РО3-РО5**
2.1	Тема 2.1. Анализ временных рядов	24	12	3	-	9	12	РО3-РО5
2.2	Тема 2.2. Ассоциативные правила	24	12	3	-	9	12	РО3-РО5
2.3	Тема 2.3. Обработка изображений	24	12	3	-	9	12	РО3-РО5
2.4	Тема 2.4. Обработка текстовых данных	24	12	3	-	9	12	РО3-РО5
2.5	Тема 2.5. Обучение с подкреплением	24	12	3	-	9	12	РО3-РО5
2.6	Тема 2.6. Рекомендательные системы	24	12	3	-	9	12	РО3-РО5
3	**Инженерия машинного обучения**	**54**	**26**	**8**	**-**	**18**	**28**	**РО6-РО8**
3.1	Тема 3.1. Жизненный цикл проектов машинного обучения	18	10	4		6	8	РО6

№ п/п	Наименование модулей (дисциплин)	Общая трудоем-кость, часов	Всего контактных часов	Контактные часы			СРС, часов	Результаты обучения
				Лекции	Лабора-торные работы	Практ. и семинарские занятия		
3.2	Тема 3.2. Управление проектом машинного обучения	18	8	2		6	10	РО6
3.3	Тема 3.3. Внедрение моделей машинного обучения	18	8	2	-	6	10	РО7-РО8
4	**Итоговая аттестация**	**36**	**6**	**-**	**-**	**6**	**30**	**РО1-РО9**
	Итого	**342**	**158**	**44**	**-**	**114**	**184**	

Календарный учебный график*
дополнительной профессиональной программы профессиональной переподготовки
«Программист - исследователь больших данных»

Наименование модулей (дисциплин)	Неделя	Объем учебной нагрузки, часов	Виды занятий (количество часов)							
			Лекция	Практ. и семинарские занятия	Лаб. работа	СРС	Консуль-тация	Контр. работа	Тест	Итоговый контроль
Методы анализа данных	1–4	108	18	36	-	54	-	-	-	Зачет
Теория вычислительного обучения	5–9	144	18	54	-	72	-	-	-	Экзамен
Инженерия машинного обучения	10–11	54	8	18	-	28	-	-	-	Зачет
Итоговая аттестация	12	36	-	6	-	30	-	-	-	Защита итоговой аттестационной работы (проекта)

I. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1.1. Аннотация программы

Отрасль информационных технологий является одной из наиболее динамично развивающихся отраслей как в мире, так и в России. Этапы качественного развития большинства отраслей экономики связаны с внедрением информационных технологий. Цифровизация широкого круга отраслей деятельности человека привела к тому, что на сегодняшний день генерируются огромные массивы структурированных и неструктурированных данных различной природы, которые необходимо не только хранить, но и извлекать знания, полезные для управления и улучшения качества жизни общества. Кроме этого, для обеспечения технологического суверенитета государства требуется большое число специалистов в области информационных технологий, среди которых немаловажную роль играют специалисты по анализу данных.

Дополнительная профессиональная программа профессиональной переподготовки «Программист - исследователь больших данных» направлена на формирование комплекса профессиональных знаний и умений, необходимых для работы в области анализа больших данных: теоретические и прикладные основы анализа больших данных, методы и инструментальные средства анализа больших данных.

По окончании обучения слушатели будут способны анализировать большие объемы данных, извлекать из них ценную информацию и делать обоснованные выводы, использовать алгоритмы машинного обучения для решения прикладных задач, разрабатывать и внедрять системы для сбора, хранения, анализа и визуализации данных, планировать и управлять аналитическими проектами, взаимодействовать с заинтересованными сторонами, четко и убедительно представлять результаты аналитической работы.

Нормативно-правовую основу разработки программы составляют:

- Федеральный закон «Об образовании в Российской Федерации» от 29.12.2012 года № 273-ФЗ;
- Методические рекомендации по разработке основных профессиональных образовательных программ и дополнительных профессиональных программ на основе профессиональных стандартов (письмо Минобрнауки РФ от 22 января 2015 г. № ДЛ-1/05);
- Профессиональный стандарт 06.042 «Специалист по большим данным» (утвержден приказом Министерства труда и социальной защиты РФ от 6 июля 2020 г. N 405н);
- Положение о дополнительном образовании и профессиональном обучении в ФГАОУ ВО «Сибирский федеральный университет», утвержденное ректором 01.04.2022 г.;
- Устав ФГАОУ ВО «Сибирский федеральный университет».

1.2. Цель программы

Цель программы – формирование компетенций, необходимых для выполнения нового вида профессиональной деятельности в области информационных технологий, а именно создание и обучение моделей и алгоритмов машинного обучения для работы с большими данными.

Программа направлена на формирование компетенций в соответствии с трудовыми функциями профессионального стандарта 06.042 «Специалист по большим данным».

Слушатель, успешно завершивший обучение по данной программе, получает диплом о профессиональной переподготовке с правом ведения нового вида профессиональной деятельности в области информационных технологий, а именно разработка и применение методов машинного обучения для решения научных и практических задач, а также приобретение по итогам прохождения Программы новой квалификации «Инженер по машинному обучению».

Программа является преемственной к основным образовательным программам высшего образования бакалавриата направлений подготовки 09.03.04 «Программная инженерия».

1.3. Характеристика нового вида профессиональной деятельности, новой квалификации

1.Область профессиональной деятельности слушателя, прошедшего обучение по программе профессиональной переподготовки, в которой может осуществлять профессиональную деятельность: разработка и применение методов машинного обучения для решения прикладных задач.

Выпускники могут осуществлять профессиональную деятельность в других областях и (или) сферах профессиональной деятельности при условии соответствия уровня их образования и полученных компетенций требованиям к квалификации работника.

2. Объекты профессиональной деятельности: методы машинного обучения. Виды профессиональной деятельности: создание и применение технологий больших данных.

3. Уровень квалификации

В соответствии с приказом Министерства труда и социальной защиты Российской Федерации от 6 июля 2020 г. «Об утверждении Профессионального стандарта «Специалист по большим данным», дополнительная профессиональная программа профессиональной переподготовки «Программист - исследователь больших данных» обеспечивает достижение 6 уровня квалификации.

1.4. Компетенции (трудовые функции) в соответствии с профессиональным стандартом (формирование новых или совершенствование имеющихся)

В соответствии с профессиональным стандартом 06.042 «Специалист по большим данным» (утв. приказом Министерства труда и социальной защиты РФ от 6 июля 2020 г. N 405н) можно выделить следующие трудовые функции, на формирование и совершенствование которых направлена Программа:

- A/03.6 Подготовка данных для проведения аналитических работ по исследованию больших данных.
- A/04.6 Проведение аналитического исследования с применением технологий больших данных в соответствии с требованиями заказчика.

1.5. Планируемые результаты обучения

В результате освоения программы слушатели будут способны:

РО1. Оценивать соответствие наборов данных задачам анализа больших данных.

РО2. Производить очистку данных для проведения аналитических работ.

РО3. Решать задачи классификации, кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных.

РО4. Проводить сравнительный анализ методов анализа больших данных.

РО5. Разрабатывать и оценивать модели больших данных.

РО6. Планировать аналитические работы с использованием технологий больших данных.

РО7. Адаптировать и развертывать модели в предметной среде.

РО8. Формировать предложения по использованию результатов анализа.

РО9. Оформлять результаты аналитического исследования для представления заказчику.

1.6. Категория слушателей

Лица, получающие высшее образование по основным профессиональным образовательным программам бакалавриата, специалитета, а также магистратуры, отнесенным к ИТ-сфере.

1.7. Требования к уровню подготовки поступающего на обучение

В соответствии с требованиями к образованию и обучению, предъявляемыми к 6 уровню квалификации профессионального стандарта 06.042 «Специалист по большим данным», необходимо иметь высшее образование или осваивать его в момент обучения на данной программе.

1.8. Продолжительность обучения

Продолжительность обучения по программе составляет 342 часа.

1.9. Форма обучения

Очно-заочная (с использованием электронного обучения и дистанционных образовательных технологий).

1.10. Требования к материально-техническому обеспечению, необходимому для реализации дополнительной профессиональной программы профессиональной переподготовки (требования к аудитории, компьютерному классу, программному обеспечению)

Обучение по программе производится в дистанционном формате с использованием сервисов вебинаров и видеоконференций.

Слушателям необходимо стандартное программное обеспечение (операционная система, офисные программы), выход в интернет, а также специализированное программное обеспечение, используемое для разработки программного обеспечения: Anaconda, PyCharm Community.

1.11. Особенности (принципы) построения дополнительной профессиональной программы профессиональной переподготовки

Особенности построения программы переподготовки «Программист - исследователь больших данных»:

– в основу проектирования программы положен компетентностный подход;

– выполнение учебных заданий, требующих практического применения знаний и умений, полученных в ходе изучения логически связанных дисциплин;

– выполнение итоговых аттестационных работ по реальному заданию;

– использование информационных и коммуникационных технологий, в том числе современных систем технологической поддержки процесса обучения, обеспечивающих комфортные условия для обучающихся, преподавателей;

– применение электронных образовательных ресурсов (дистанционное, электронное, комбинированное обучение и пр.).

1.12. Документ об образовании: диплом о переподготовке установленного образца.

II. ОЦЕНКА КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ

2.1. Формы аттестации, оценочные материалы, методические материалы

Программа предусматривает проведение текущей и итоговой аттестации. Текущая аттестация слушателей проводится по дисциплинам на основе выполнения практических заданий. Промежуточная аттестация осуществляется путем сдачи зачетов и экзаменов по соответствующим дисциплинам.

Итоговая аттестационная работа выполняется индивидуально в форме проектной работы.

Итоговой аттестационной работой является защита итоговой аттестационной работы, которая проходит в синхронном формате.

2.2. Требования и содержание итоговой аттестации

К итоговой аттестации допускаются слушатели, выполнившие учебный план программы профессиональной переподготовки, самостоятельные задания в каждой дисциплине в полном объеме за все время обучения.

Итоговая аттестация по программе включает защиту итоговой аттестационной работы в форме проекта, которая проходит в синхронном формате.

Основная цель итоговой аттестационной работы (ИАР) – выполнить работу, демонстрирующую уровень освоения теоретического и практического материала программы, а также подготовленности к самостоятельной профессиональной деятельности.

ИАР выполняется индивидуально. Защита ИАР включает презентацию работы, вопросы по различным разделам программы. Защита ИАР дает возможность продемонстрировать уровень приобретенных слушателем профессиональных компетенций.

Слушатель предоставляет результат выполненной работы в формате PDF. Объем презентации следует выбирать исходя из длительности выступления (обычно – не более 5-7 минут). В выступлении должны быть четко обозначены тема, область и актуальность работы, постановка цели и задач, приведены результаты, полученные слушателем и проведен их анализ.

Требования к содержанию пояснительной записки ИАР

1. Введение. Описание задачи.
2. Постановка задачи машинного обучения.
3. Описание набора данных и шагов по его подготовке для обучения модели.
4. Описание архитектуры модели (моделей) и метода обучения.
5. Оценка качества обучения, графики.
6. Результаты работы моделей.
7. Заключение. Описание достигнутых результатов, если использовалось несколько моделей, сравнение качества их работы.
8. Список использованных источников.
9. Приложение. Полный код решения.

Требования к устному докладу

1. Приветствие, обращение к членам комиссии и представление.
2. Тема итоговой аттестационной работы.
3. Актуальность, цель и задачи работы.
4. Набор данных и его подготовка для обучения модели.
5. Исследуемые модели и алгоритмы решения поставленной задачи.
6. Анализ результатов работы.
7. Заключение.

Продолжительность выступления – 7-8 минут.

Критерии оценивания итоговой аттестационной работы

Критерий	Показатели выполнения	Баллы (мин/макс)
Содержание работы	Обоснована актуальность работы	0/1
	Цели и задачи итоговой аттестационной работы определены и согласованы между собой	0/1
	Показана практическая значимость работы	0/1
	Проведен анализ имеющихся аналогов	0/1
	Обоснован выбор средства реализации требований к компьютерному программному обеспечению	0/1
	Представлены результаты сравнительного анализа методов анализа больших данных	0/1
	Заключение работы содержит оценку результативности и перспектив результатов работы	0/1
Доклад/защита работы	Выступление соответствует требованиям публичной речи: материал изложен точно, доступно	0/1
	Презентация оформлена в деловом стиле. Информация представлена в виде тезисов, цитат	0/1
	Получены ответы на вопросы, заданные членами аттестационной комиссии	0/1
Всего		10 баллов

- Оценка «отлично» ставится, если слушатель набрал 9–10 баллов.
- Оценка «хорошо» ставится, если слушатель набрал 7–8 баллов.
- Оценка «удовлетворительно» ставится, если слушатель набрал 5–6 баллов.

Итоговая аттестационная работа защищается в синхронном формате перед аттестационной комиссией; работа представляется с помощью устного доклада и демонстрации презентации.

Защита итоговой аттестационной работы является обязательной.

По результатам защиты ИАР аттестационная комиссия принимает решение о присвоении слушателям по результатам освоения дополнительной профессиональной программы профессиональной переподготовки квалификации «Инженер по машинному обучению», о предоставлении права заниматься профессиональной деятельностью в сфере создания и применения технологий больших данных и выдаче диплома о профессиональной переподготовке.

III. ОСНОВНОЕ СОДЕРЖАНИЕ ПРОГРАММЫ

3.1. План учебной деятельности

Результаты обучения	Учебные действия/ формы текущего контроля	Используемые ресурсы/ инструменты/технологии
РО1. Оценивать соответствие наборов данных задачам анализа больших данных	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО2. Производить очистку данных для проведения аналитических работ	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО3. Решать задачи классификации, кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных.	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий и итоговой работы. Форма текущего контроля: Зачет/Экзамен	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО4. Проводить сравнительный анализ методов анализа больших данных.	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий и итоговой работы. Форма текущего контроля: Зачет/Экзамен	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО5. Разрабатывать и оценивать модели больших данных.	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО6. Планировать аналитические работы с использованием технологий больших данных.	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО7. Адаптировать и развертывать модели в предметной среде.	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО8. Формировать предложения по использованию результатов анализа.	Лекция, изучение основной и дополнительной литературы, выполнение практических заданий	Материалы в системе электронного обучения . Системы видеоконференцсвязи
РО9. Оформлять результаты аналитического исследования для представления заказчику	Выполнение итоговой работы	- Материалы в системе электронного обучения . Системы видеоконференцсвязи

3.2. Виды и содержание самостоятельной работы

Самостоятельная работа слушателя (СРС) предполагает углубление и закрепление теоретических знаний. СРС включает следующие виды самостоятельной деятельности: самостоятельное углубленное изучение дополнительной литературы по темам дисциплин, выполнение индивидуальных заданий, работа по проекту. Выполнение СРС предполагается в дистанционном режиме.

РАБОЧАЯ ПРОГРАММА
модуля (дисциплины)
«Методы анализа данных»

1. Аннотация

В рамках освоения данной дисциплины слушатели познакомятся с основными принципами и методами анализа данных, научатся применять изученные методы анализа данных при решении реальных практических задач в различных сферах практической деятельности, овладеют навыками разработки инструментальных средств анализа данных.

Цель модуля (результаты обучения)

По окончанию обучения на данной дисциплине слушатели будут способны:

РО1. Оценивать соответствие наборов данных задачам анализа больших данных.

РО2. Производить очистку данных для проведения аналитических работ.

РО3. Решать задачи классификации, кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных.

РО4. Проводить сравнительный анализ методов анализа больших данных.

РО5. Разрабатывать и оценивать модели больших данных.

2. Содержание

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
Тема 1.1. Введение в анализ данных	Введение в машинное обучение. Обучение моделей «с учителем» и «без учителя». Обучающее и тестовое множество. Эффект переобучения и недообучения. Смещение и разброс. Функция потерь и функционал качества. Метод перекрестной	-	Изучение дополнительного материала по темам: Введение в машинное обучение. Обучение моделей «с учителем» и «без учителя». Обучающее и тестовое множество. Эффект переобучения и недообучения. Смещение и разброс. Функция потерь и функционал качества. Метод перекрестной проверки. Основные этапы проектов машинного обучение. Особенности накопления выборки данных и

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
	проверки. Основные этапы проектов машинного обучение. Особенности накопления выборки данных и количественная его оценка. Разметка данных (2 ч.)		количественная его оценка. Разметка данных. Оценка качества работы моделей (2 ч.)
Тема 1.2. Первичная обработка данных	Ключевые задачи в подготовке данных. Восстановление пропущенных значений. Обнаружение выбросов. Балансировка данных. Обработка категориальных признаков. Масштабирование признаков. Визуализация выборочных данных. Описательная статистика для переменных. (2 ч.)	Предварительная обработка и разведочный анализ данных на языке Python (4 ч.). *Задание 1*. Проверка правдоподобности исходных данных. Поиск аномальных значений. Поиск и восстановление пропущенных значений. Преобразование данных. Визуализация данных	Изучение дополнительного материала по темам: Ключевые задачи в подготовке данных. Восстановление пропущенных значений. Обнаружение выбросов. Балансировка данных. Обработка категориальных признаков. Масштабирование признаков. Визуализация выборочных данных. Описательная статистика для переменных. (8 ч.)
Тема 1.3. Задача классификации	Постановка задачи классификации. Линейные модели классификации. Байесовская теория принятия решений. Логистическая регрессия.	Классификация данных на языке Python (8 ч.) *Задание 2*. Обучение моделей классификации. Подбор оптимальных параметров моделей. Оценка качества	Изучение дополнительного материала по темам: Постановка задачи классификации. Линейные модели классификации. Байесовская теория принятия решений. Логистическая регрессия. Метод ближайших соседей. Метрики оценки

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
	Метод ближайших соседей. Метрики оценки качества в задаче классификации (4 ч.)	построенных моделей.	качества в задаче классификации (10 ч.)
Тема 1.4. Задача восстановления регрессии	Постановка задачи восстановления регрессии. Простая линейная регрессия. Множественная линейная регрессия. Метод наименьших квадратов. Метрики оценки качества в задаче восстановления регрессии. Нелинейная регрессия (4 ч.)	Выполнение регрессионного анализа на языке Python (8 ч.). *Задание 3*. Построение регрессионных моделей. Подбор оптимальных параметров моделей. Оценка качества построенных моделей	Изучение дополнительного материала по темам: Постановка задачи восстановления регрессии. Простая линейная регрессия. Множественная линейная регрессия. Метод наименьших квадратов. Метрики оценки качества в задаче восстановления регрессии. Нелинейная регрессия (10 ч.)
Тема 1.5. Деревья решений	Введение в деревья решений. Алгоритмы построения деревьев решений. Редукция решающих деревьев (2 ч.)	Построения деревьев решений на языке Python (6 ч.). *Задание 4*. Построение моделей на основе деревьев решений для решения задачи классификации. Подбор оптимальных параметров моделей. Оценка качества построенных моделей	Изучение дополнительного материала по темам: Введение в деревья решений. Алгоритмы построения деревьев решений. Редукция решающих деревьев (8 ч.)

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
Тема 1.6. Ансамбли моделей	Ансамблевые методы: Бэггинг, бустинг, стекинг (2 ч.)	Построения ансамблей алгоритмов на языке Python (6 ч.). *Задание 5*. Построение ансамблей алгоритмов (беггинг, бустинг и стекинг) для решения задачи классификации. Подбор оптимальных параметров моделей. Оценка качества построенных моделей	Изучение дополнительного материала по темам: Ансамбли моделей (8 ч.)
Тема 1.7. Кластеризация	Постановка задачи кластеризации. Графовые методы кластеризации. Иерархические методы кластеризации. Статистические методы кластеризации. Методы на основе плотности (2 ч.)	Выполнение кластерного анализа на языке Python (4 ч.). *Задание 6*. Построение моделей кластеризации данных. Подбор оптимальных параметров моделей. Оценка качества построенных моделей	Изучение дополнительного материала по темам: Постановка задачи кластеризации. Графовые методы кластеризации. Иерархические методы кластеризации. Статистические методы кластеризации. Методы на основе плотности (8 ч.)

3. Условия реализации программы модуля

Организационно-педагогические условия реализации программы

Обучение по программе реализовано в формате смешанного обучения, с применением активных технологий совместного обучения в электронной среде (синхронные и асинхронные занятия). Лекционный материал представляется в виде синхронных лекций, текстовых материалов и презентаций. Данные

материалы сопровождаются практическими заданиями. Изучение теоретического материала (СРС) предполагается до и после синхронной части работы.

Материально-технические условия реализации программы

Синхронные занятия реализуются на базе инструментов видеоконференцсвязи и включают в себя лекционные и практические занятия. Для проведения синхронных занятий (вебинаров) применяется программа видеоконференцсвязи. При проведении лекций, практических занятий, самостоятельной работы слушателей используется следующее оборудование: компьютер с наушниками или аудиоколонками, микрофоном и веб-камерой. Программное обеспечение (обновленное до последней версии): браузер Google Chrome, Anaconda, PyCharm Community.

Учебно-методическое и информационное обеспечение программы

Дисциплина может быть реализована как очно, так и заочно, в том числе, с применением дистанционных образовательных технологий. Она включает занятия лекционного типа, интерактивные формы обучения, семинарские, практические занятия.

Содержание комплекта учебно-методических материалов

По данной дисциплине имеется учебно-методический комплекс (УМК), который содержит: систему навигации по дисциплине (учебно-тематический план, график работы по дисциплине, сведения о результатах обучения), текстовые материалы к лекциям, практические задания, списки основной и дополнительной литературы.

Литература

1. Бурков А. Машинное обучение без лишних слов. Библиотека программиста. – М., 2020. – 192 с.

2. Грас Д. Data Science. Наука о данных с нуля: пер. с англ. – СПб.: БХВ-Петербург, 2017. – 336 с.

3. Джоши П. Искусственный интеллект с примерами на Python. – М., 2019. – 448 с.

4. Маккинли У. Python и анализ данных. – Саратов: Профобразование, 2019. – 482 с.

5. Мирджалили В., Рашка С. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow. – М., 2020. – 848 с.

6. Мыльников Л.А. Статистические методы интеллектуального анализа данных. – СПб.: БХВ-Петербург, 2021 – 240 с.

7. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. – М., 2017. – 480 с.

8. Плас Дж. В.Python для сложных задач: наука о данных и машинное обучение. – СПб.: Петербург, 2018 – 576 с.

9. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data, Python и наука о данных. – М., 2017. – 336 с.

10. Скиена С. Наука о данных: учебный курс. Пер. с англ. – СПб.: ООО «Диалектика», 2020. – 544 с.

11. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных: пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.

12. Хасти Т., Тибришани Р. Основы статистического обучения: интеллектуальный анализ данных, логический вывод и прогнозирование. – М., 2020. – 768 с.

13. Элбон К. Машинное обучение с использованием Python. Сборник рецептов. – СПб.: БХВ-Петербург, 2020. – 384 с.

4. Оценка качества освоения программы дисциплины

Форма аттестации по дисциплине — зачет. Зачет выставляется за выполненные практические задания.

Перечень заданий

Практические задания дисциплины

Практическая работа № 1. Предварительная обработка данных.

Цель: знакомство с основными задачами предварительной обработки исходных данных, изучение основных методов предварительной обработки данных, формирование навыков выполнения предварительной обработки исходных данных с помощью языка программирования Python.

Задачи

Выполнение практической работы предполагает решение следующий задач:

1. Визуальный анализ исходных данных
2. Поиск аномальных значений
3. Поиск и восстановление отсутствующих значений
4. Преобразование данных

Общая последовательность действий

1. Визуальный анализ данных

Построить визуальное представление для каждого столбца (признака) в исходном наборе данных. Провести анализ полученных графиков.

2. Провести проверку правдоподобности исходных данных. Проверка правдоподобности исходных данных должна включать проверку типов исходных данных, лишних пропусков, невозможных значений и т.п. Привести найденные значения к нужному формату.

3. Поиск аномальных значений. Провести поиск значений в исходном наборе данных, резко отличающихся от других значений (выбросов). Строки с найденными выбросами удалить из исходного набора данных.

Провести анализ полученных результатов. Использовать результаты очистки данных, полученных с помощью метода сигм.

4. Поиск и восстановление пропущенных значений. Провести поиск пропущенных значений в исходных данных. Вывести статистику по пропускам для каждого признака. Восстановить пропущенные значения.

5. Преобразование данных. Привести числовые признаки к стандартному виду. Для категориальных признаков выполнить их кодировку.

Практическая работа № 2. Задача классификации

Цель: знакомство с теоретическими основами задачи классификации объектов, формирование навыков решения задачи классификации с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Предварительная обработка исходных данных
2. Обучение базовых моделей классификации
3. Подбор оптимальных параметров моделей классификации
4. Оценка качества построенных моделей на тестовой выборке

Общая последовательность действий

1. Загрузить данные для обучения и для теста.
2. Выполнить предварительную обработку исходных данных (в случае необходимости)
3. Построить модели классификаторов с параметрами, подобранными на перекрестной проверке (cross validation).
4. Предсказать целевую переменную для тестовой выборки.
5. Сравнить полученные результаты по метрикам accuracy, precision, recall, f1- score. Выделить наилучшую модель. Построить ROC-кривую и вычислить ROC-AUC на тренировочной и тестовой выборках. Визуализировать результаты с помощью ConfusionMatrix.

Практическая работа № 3. Восстановление регрессии

Цель: знакомство с теоретическими регрессионного анализа, формирование навыков применения регрессионного анализа для решения задачи восстановления функциональных зависимостей с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Предварительная обработку исходных данных
2. Обучение базовых регрессионных моделей
3. Подбор оптимальных параметров регрессионных моделей

4. Оценка качества построенных моделей на валидационной/тестовой выборке.

Общая последовательность действий

1. Загрузить исходные данные.
2. Выполнить предварительную обработку исходных данных (в случае необходимости)
3. Построить регрессионные модели с параметрами, подобранными на перекрестной проверке (cross validation).
4. Спрогнозировать значение выходной переменной для тестовой выборки.
6. Рассчитать значения метрик качества. Выделить наилучшую модель.

Практическая работа № 4. Деревья решений

Цель: знакомство с теоретическими основами построения деревьев решений, формирование навыков построения деревьев решений для решения задач классификации и регрессии с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Предварительная обработку исходных данных
2. Построение базовых моделей на основе деревьев решений
3. Подбор оптимальных параметров моделей
4. Оценка качества построенных моделей на валидационной/тестовой выборке

Общая последовательность действий

1. Загрузить исходные данные для обучения и для теста.
2. Выполнить предварительную обработку исходных данных (в случае необходимости)
3. Построить дерево решений. Визуализировать результат.
4. Исследовать зависимость качества прогнозирования от настраиваемых параметров (на тестовой выборке, на контрольной выборке, в ходе перекрестной проверки). Построить график зависимости значений метрик от значений параметров модели. Сделать выводы об оптимальных значениях параметров.
5. Исследовать, приводит ли обрезка дерева к улучшению результата на контрольной выборке.
6. Сравнить результаты, полученные с использованием единичного дерева с результатами, полученными во второй и третьей практической работе.

Практическая работа № 5. Ансамбли моделей

Цель: знакомство с теоретическими основами построения ансамблей алгоритмов, формирование навыков построения ансамблей моделей для решения задач классификации и регрессии с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Предварительная обработку исходных данных
2. Построение ансамблей моделей (беггинг, бустинг и стекинг)
3. Подбор оптимальных параметров моделей
4. Оценка качества построенных моделей на валидационной/тестовой выборке

Общая последовательность действий

1. Загрузить исходные данные для обучения и для теста.
2. Выполнить предварительную обработку исходных данных (в случае необходимости).
3. Построить композиции алгоритмов с использованием методов бэггинга, бустинга и стекинга.
4. Исследовать зависимость качества прогнозирования от настраиваемых параметров (на тестовой выборке, на контрольной выборке, в ходе перекрестной проверки). Построить график зависимости значений метрик от значений параметров модели. Сделать выводы об оптимальных значениях параметров. Для технологии стекинга подобрать оптимальный состав базовых алгоритмов.
5. Сравнить результаты, полученные с использованием ансамблей результатами, полученными с использованием единичного дерева из практической работы №4 и с помощью реализованных алгоритмов во второй и третьей практических работах.

Практическая работа № 6. Кластеризация

Цель: знакомство с теоретическими основами задачи кластеризации, формирование навыков решения задачи кластеризации с помощью языка программирования Python.

Задачи:

1. Предварительная обработка исходных данных
2. Обучение базовых моделей кластеризации
3. Подбор оптимальных параметров моделей
4. Оценка качества построенных моделей

Общая последовательность действий:

1. Используя данные по задаче классификации (предварительно убрав столбец с метками классов), выполнить кластеризацию наблюдений с использованием метода k-means со значением k, равным числу классов. Оценить, насколько хорошо полученные классы согласуются с истинными метками классов.
2. Исследовать качество кластеризации методом k-means, варьируя значения настраиваемых параметров (при различном количестве кластеров и различных мерах расстояния).
3. Написать выводы относительно получаемых результатов.

Задания для самостоятельной работы

В самостоятельные работы входит изучение дополнительного материала по темам дисциплины и закрепление заданий с практических занятий.

Критерии оценивания заданий и/или контрольных вопросов

Баллы	3 балла	4 балла	5 балла
Критерий	Задание выполнено частично, требует серьезной доработки	Задание выполнено, но требует некоторой доработки	Задание выполнено полностью, не требует доработки

РАБОЧАЯ ПРОГРАММА
дисциплины
«Теория вычислительного обучения»

1. Аннотация

В рамках освоения данной дисциплины слушатели познакомятся с основными принципами и методами в области теории вычислительного обучения, научатся применять изученные методы анализа данных при решении реальных практических задач в различных сферах практической деятельности, овладеют навыками разработки инструментальных средств анализа данных.

Цель дисциплины (результаты обучения)

По окончанию обучения на данной дисциплине слушатели будут способны:

РО3. Решать задачи классификации, кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных.

РО4. Проводить сравнительный анализ методов анализа больших данных.

РО5. Разрабатывать и оценивать модели больших данных.

2. Содержание

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
Тема 2.1. Анализ временных рядов	Понятие временного ряда. Классификация временных рядов, компоненты временного ряда. Задачи анализа временных рядов. Методы преобразования временных рядов. Методы прогнозирования временных рядов, стратегии получения долгосрочного прогноза (3 ч.)	Анализ временных рядов на языке Python (9 ч.) *Задание 1*. Визуализация временных рядов. Предварительная обработка. Построение моделей временных рядов	Изучение дополнительного материала по темам: Понятие временного ряда. Классификация временных рядов, компоненты временного ряда. Задачи анализа временных рядов. Методы преобразования временных рядов. Методы прогнозирования временных рядов, стратегии получения долгосрочного прогноза (12 ч.)

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
Тема 2.2. Ассоциативные правила	Введение в ассоциативные правила. Алгоритм Apriori. Алгоритм FP-Growth. Ассоциативная классификация. Секвенциальный анализ (3 ч.)	Поиск ассоциативных правил на языке Python (9 ч.) *Задание 2*. Поиск ассоциативных правил. Визуализация ассоциативных правил	Изучение дополнительного материала по темам: Введение в ассоциативные правила. Алгоритм Apriori. Алгоритм FP-Growth. Ассоциативная классификация. Секвенциальный анализ (12 ч.)
Тема 2.3. Обработка изображений	Цифровые изображения. Обработка изображений. Извлечение признаков изображений. Методы классификации изображений (3 ч.)	Обработка изображений на языке Python (9 ч.) *Задание 3*. Градационные, гистограммные и геометрические преобразования изображений. Фильтрация изображений. Выделение контуров на изображении. Бинаризация изображений.	Изучение дополнительного материала по темам: Цифровые изображения. Обработка изображений. Извлечение признаков изображений. Методы классификации изображений (12 ч.)
Тема 2.4. Обработка текстовых данных	Предобработка текстовых данных. Методы извлечения признаков из текстовых данных. Задача классификации в текстовом анализе. Тематическое моделирование. Извлечении	Обработка текстовых данных на языке Python (9 ч.) *Задание 4*. Предварительная обработка текстовых данных. Векторизация текстовых данных.	Изучение дополнительного материала по темам: Предобработка текстовых данных. Методы извлечения признаков из текстовых данных. Задача классификации в текстовом анализе. Тематическое моделирование. Извлечении именованных сущностей (12 ч.)

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
	именованных сущностей (3 ч.)	Классификация текстовых данных	
Тема 2.5. Обучение с подкреплением	Постановка задачи обучения с подкреплением, элементы системы обучения с подкреплением. Многорукие бандиты. Марковский процесс принятия решений. Динамическое программирование. Методы временных различий (3 ч.)	Обучение с подкреплением на языке Python (9 ч.) *Задание 5*. Построение системы обучения с подкреплением	Изучение дополнительного материала по темам: Постановка задачи обучения с подкреплением, элементы системы обучения с подкреплением. Многорукие бандиты. Марковский процесс принятия решений. Динамическое программирование. Методы временных различий (12 ч.)
Тема 2.6. Рекомендательные системы	Рекомендательные системы. Методы коллаборативной фильтрации. Факторизационные машины. Контентная фильтрация. Рекомендательные системы, основанные на знаниях. Метрики оценки качества рекомендательных систем (3 ч.)	Построение рекомендательных систем на языке Python (9 ч.) *Задание 6*. Построение рекомендательных систем. Формирование рекомендаций	Изучение дополнительного материала по темам: Рекомендательные системы. Методы коллаборативной фильтрации. Факторизационные машины. Контентная фильтрация. Рекомендательные системы, основанные на знаниях. Метрики оценки качества рекомендательных систем (12 ч.)

3. Условия реализации программы модуля

Организационно-педагогические условия реализации программы

Обучение по программе реализовано в формате смешанного обучения, с применением активных технологий совместного обучения в электронной среде (синхронные и асинхронные занятия). Лекционный материал представляется в

виде синхронных лекций, текстовых материалов и презентаций. Данные материалы сопровождаются практическими заданиями. Изучение теоретического материала (СРС) предполагается до и после синхронной части работы.

Материально-технические условия реализации программы

Синхронные занятия реализуются на базе инструментов видеоконференцсвязи и включают в себя лекционные и практические занятия. Для проведения синхронных занятий (вебинаров) применяется программа видеоконференцсвязи. При проведении лекций, практических занятий, самостоятельной работы слушателей используется следующее оборудование: компьютер с наушниками или аудиоколонками, микрофоном и веб-камерой. Программное обеспечение (обновленное до последней версии): браузер Google Chrome, Anaconda, PyCharm Community.

Учебно-методическое и информационное обеспечение программы

Дисциплина может быть реализована как очно, так и заочно, в том числе, с применением дистанционных образовательных технологий. Она включает занятия лекционного типа, интерактивные формы обучения, семинарские, практические занятия.

Содержание комплекта учебно-методических материалов

По данной дисциплине имеется учебно-методический комплекс (УМК), который содержит: систему навигации по дисциплине (учебно-тематический план, график работы по дисциплине, сведения о результатах обучения), текстовые материалы к лекциям, практические задания, списки основной и дополнительной литературы.

Литература

1. Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – СПб.: Питер, 2019. - 368 с.

2. Саттон Р. С., Барто Э. Дж. С21 Обучение с подкреплением: Введение. 2-е изд. / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2020. – 552 с.

3. Лонца А. Алгоритмы обучения с подкреплением на Python / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2020. – 286 с.

4. Нильсен Э. Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение. : Пер. с англ. – СПб. : ООО “Диалектика”, 2021. – 544 с.

5. Солем Я. Э. Программирование компьютерного зрения на языке Python / пер. с англ. Слинкин А.А. – М.: ДМК Пресс, 2016. – 312 с.

6. Клетте Р. Компьютерное зрение. Теория и алгоритмы / пер. с англ. Слинкин А.А. – М.: ДМК Пресс, 2019. – 506 с.

7. Ким Фальк Рекомендательные системы на практике / пер. с англ. Д. М. Павлова. – М.: ДМК Пресс, 2020. – 448 с.

8. Селянкин В. В. Компьютерное зрение. Анализ и обра. ботка изображений: учебное пособие для вузов / Селянкин В. В. – Санкт-Петербург: Лань, 2023. - 152 с.

4. Оценка качества освоения программы модуля (формы аттестации, оценочные и методические материалы)

Форма аттестации по дисциплине — экзамен.

Результаты экзамена определяются оценками "отлично", "хорошо", "удовлетворительно", "неудовлетворительно".

Оценка "отлично" ставится за свободное владение материалом, аргументированные ответы на основные и дополнительные вопросы, знание понятийного аппарата по теме вопроса.

Оценка "хорошо" ставится за полные и аргументированные ответы на основные и дополнительные вопросы, знание понятийного аппарата по теме вопроса при незначительных упущениях и неточностях.

Оценка "удовлетворительно" может быть выставлена при неполных и слабо аргументированных ответах только в том случае, если экзаменуемый обнаруживает понимание существа поставленных в билете вопросов, владеет понятийным аппаратом, т.е. владеет программным материалом в объеме, необходимом для дальнейшей учебы и работы.

Список вопросов по дисциплине:

1. Понятие временного ряда. Основные задачи анализа временных рядов. Составляющие временного ряда. Методы выявления тренда
2. Стационарный временной ряд. Тесты на стационарность. Методы приведения к стационарному временному ряду.
3. Автокорреляционная и частная автокорреляционная функция.
4. Сглаживание временных рядов.
5. Модели стационарных временных рядов: процессы авторегрессии и процессы скользящего среднего
6. Смешанные процессы авторегрессии
7. Построение прогнозов для временных рядов: методы долгосрочного прогнозирования. Модели экспоненциального сглаживания.
8. Поиск ассоциативных правил. Алгоритм Apriori.
9. Поиск ассоциативных правил. Алгоритм FP-Growth.
10. Цифровое изображение. Градационные и гистограммные преобразования изображений.
11. Цифровое изображение. Геометрические преобразования изображений.
12. Бинаризация изображений.
13. Модели шума. Линейные низкочастотные фильтры. Нелинейные фильтры.
14. Задача выделения контуров. Высокочастотные линейные фильтры. Метод Кэнни.
15. Морфологическая обработка изображений.
16. Глобальные признаки изображений (цвет, форма, текстура).
17. Детекторы и дескрипторы признаков. Требования к ключевым точкам.
18. Детекторы углов.
19. Метод SIFT.

20. Метод HOG.
21. Основные операции предобработки текстовых данных.
22. Методы векторизации текста.
23. Тематическое моделирование. Латентный семантический анализ.
24. Анализ тональности текста.
25. Обучение с подкреплением. Элементы системы обучения с подкреплением.
26. Задача многорукого бандита.
27. Марковский процесс принятия решений.
28. TD-обучение. Q-learning.
29. Рекомендательные системы. Коллаборативная фильтрация.
30. Модели со скрытыми переменными. Факторизационные машины.
31. Контентная фильтрация.
32. Метрики оценки качества рекомендательных систем.

Перечень заданий

Практические задания дисциплины

Практическая работа №1 Анализ временных рядов

Цель: знакомство с основными методами и подходами анализа временных рядов, формирование навыков построения прогноза временного ряда на языке Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Визуализация исходных данных.
2. Предварительный обработка исходных данных
3. Описание исходных данных
4. Анализ динамики временного ряда
5. Проверка на стационарность
6. Исследование тенденции временного ряда
7. Построение моделей прогнозирования значений временного ряда
8. Оценка качества построенных моделей на тестовой выборке

Общая последовательность действий

1. Ознакомиться со своим вариантом
2. Построить график временного ряда
3. Выполнить предварительную обработку исходных данных (в случае необходимости)
4. Выполнить проверку временного ряда на стационарность (с помощью ACF и PACF и теста Дики-Фуллера). Дать текстовую интерпретацию полученным результатам

5. Выполнить расчет описательных статистик. Дать графическую и текстовую интерпретацию полученным результатам
6. Исследовать тенденцию временного ряда (проверить гипотезу о наличии тенденции во временном ряду, оценить параметры кривой роста, проверить адекватность трендовой модели)
7. Построить модели авторегрессии и скользящего среднего. Проверить адекватность моделей
8. Построить краткосрочный и долгосрочные прогнозы рассматриваемой величины

Практическая работа №2 Поиск ассоциативных правил

Цель: знакомство с методами поиска ассоциативных правил, формирование навыков выполнения поиска ассоциативных правил на языке Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Подготовка исходных данных
2. Генерация частных наборов данных
3. Генерация ассоциативных правил

Общая последовательность действий

1. Выбрать страну (согласно варианту) из исходных данных для анализа рыночной корзины.
2. Выполнить предварительную обработку данных
3. Представить исходный набор данных в виде множества транзакций (использовать one-hot encoding)
4. Сгенерировать частные наборы элементов и ассоциативные правила
5. Для ассоциативных правил вычислить дополнительные меры $lift$, $leverage$, $conviction$
6. Исследовать влияние величины пороговых значений для мер $support$ и $confidence$ на число генерируемых правил. Построить соответствующие графики.
7. Визуализировать полученные правила
8. Сравнить результаты с покупками в другой стране (согласно варианту)
9. Сделать выводы по результатам проведенных исследований

Практическая работа № 3. Разработка рекомендательных систем

Цель: знакомство с основными методами и подходами реализации рекомендательных систем, формирование навыков построения рекомендательных систем на языке Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Визуализация исходных данных.
2. Предварительный обработка исходных данных
3. Реализация рекомендательной системы
4. Оценка качества формируемых рекомендаций с помощью специальных метрик

Общая последовательность действий

1. Ознакомиться с выбранными данными
2. Выполнить предварительную обработку исходных данных (в случае необходимости)
3. Построить рекомендательную систему с использованием гибридного подхода.
4. Построить нейросетевую рекомендательную систему
5. Сформировать рекомендации с помощью реализованных подходов.
6. Сравнить полученные рекомендации с помощью специальных метрик.

Практическая работа № 4. Обучение с подкреплением

Цель: знакомство с основными методами и подходами обучения с подкреплением, формирование навыков реализации алгоритмов обучения с подкреплением на языке Python.

Задачи:

Выполнение практической работы предполагает решение задачи о многоруком бандите.

Общая последовательность действий

1. Решить задачу о многоруком бандите, используя алгоритмы EpsGreedy, Softmax и UCB1.
2. Подобрать параметры алгоритмов.
3. Построить графики среднего выигрыша.

Практическая работа № 5. Обработка текстовых данных

Цель: знакомство с методами анализа текстовых данных в рамках решения задачи анализа текстовых данных, формирование навыков выполнения анализа данных на языке Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Подготовка исходных данных

2. Обучение базовых моделей классификации
3. Оценка качества построенных моделей на тестовой выборке

Общая последовательность действий

1. Выполнить предварительную обработку текстовых данных.
2. Реализовать базовые алгоритмы классификации.
3. Провести валидацию моделей на текстовых данных
4. Сравнить результаты классификации при использовании методов векторизации «мешок слов», TF-IDF и Word2Vec.
5. Реализовать подход на основе тонального словаря. Сравнить данный подход с подходами на основе машинного обучения
6. Сделать выводы по результатам проведенных исследований

Практическая работа № 6. Обработка изображений

Цель: знакомство с методами обработки изображений, формирование навыков выполнения обработки изображений на языке Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Исследование методов обработки изображений
2. Исследование детекторов и дескрипторов ключевых точек
3. Реализация моделей и алгоритмов распознавания объектов на изображении

Общая последовательность действий

1. Подготовить изображения для исследования
2. Исследовать методы обработки изображений

2.1 Обработать изображения с помощью методов геометрического преобразования (аффинные преобразования: сдвиг, поворот, изменение масштаба, проективные преобразования, отражение относительно горизонтали или вертикали). Обработать изображения с помощью методов преобразования яркости и контраста.

2.2 Обработать изображения разными методами фильтрации (низкочастотные фильтры, нелинейные фильтры). Исследовать работу фильтров для разных видов шумов (гауссовский, импульсный шум типа «соль», импульсный шум типа «перец», импульсный шум со случайным значением импульсов). Оценить качество восстановления изображения (например, с помощью метрики «пиковое отношение сигнал-шум» (peak signal-to-noise ratio, PSNR)).

2.3 Получить контурные изображения различными методами выделения контуров (высокочастотные фильтры: Робертса, Превитта, Собела, Лапласа, метод Кэнни). Исследовать качество выделения контуров в зависимости от выбора порога. Определить, при каком значении порога качество контурного

изображения будет наилучшее. Качество контурного изображения можно оценить визуально и с помощью среднеквадратичной ошибки (для вычисления критерия необходимо сравнить сформированное контурное изображение с идеальным). Исследовать качество выделения контуров от уровня шума.

3 Исследование детекторов и дескрипторов

3.1 Обработать изображения с помощью разных методов поиска ключевых точек (SIFT, SURF, BRISK, ORB). Отобразить ключевые точки на изображении. Выполнить поиск похожих объектов с помощью дескрипторов данных методов на других изображениях (изображения без и со схожими объектами). Исследовать инвариантность методов относительно преобразования исходного изображения (смещение, поворот, изменение масштаба, изменение яркости, изменение точки положения камеры).

4 Решение задачи распознавания фотографий газетных судоку, сделанных с помощью смартфона. Поле судоку представляет собой сетку размером 9 × 9. Каждая ячейка может быть пустой или содержать значение от 1 до 9.

4.1 Разработать алгоритм распознавания сетки судоку. Визуализировать результат работы алгоритмы

4.2 Разработать модель распознавания цифр в ячейках сетки судоку. В качестве данных для обучения использовать базу данных MNIST. Для извлечения признаков из данных использовать гистограммы ориентированных градиентов (HOG, SIFT, SURF). Построить модели классификаторов, оценить качество их работы.

Задания для самостоятельной работы

В самостоятельные работы входит изучение дополнительного материала по темам дисциплины и закрепление заданий с практических занятий.

Критерии оценивания заданий и/или контрольных вопросов

Баллы	3 балла	4 балла	5 балла
Критерий	Задание выполнено частично, требует серьезной доработки	Задание выполнено, но требует некоторой доработки	Задание выполнено полностью, не требует доработки

РАБОЧАЯ ПРОГРАММА
дисциплины
«Инженерия машинного обучения»

1. Аннотация

В рамках освоения данной дисциплины слушатели познакомятся с передовыми практиками и инструментами в области машинного обучения, получат практические навыки по работе с проектами машинного обучения, а также научатся оценивать и улучшать эффективность моделей машинного обучения в различных условиях. Изучение данной дисциплины позволит слушателям лучше понимать технические аспекты машинного обучения, а также они научатся эффективно внедрять эти технологии в реальные бизнес-процессы.

Цель дисциплины (результаты обучения)

По окончанию обучения на данной дисциплине слушатели будут способны:

РО6. Планировать аналитические работы с использованием технологий больших данных.

РО7. Адаптировать и развертывать модели в предметной среде.

РО8. Формировать предложения по использованию результатов анализа.

2. Содержание

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
Тема 3.1. Жизненный цикл проектов машинного обучения	Жизненный цикл проектов машинного обучения. Сбор и анализ требований, постановка задачи. (4 ч.)	(6 ч.) *Задание 1*. Сбор и анализ требований к системе машинного обучения. Постановка задачи машинного обучения и анализ решений.	Изучение дополнительного материала по темам: Жизненный цикл проектов машинного обучения. Сбор и анализ требований, постановка задачи. Анализ предметной области и аналогов (8 ч.)
Тема 3.2. Управление проектом машинного обучения	Оценка сроков и стоимости разработки. Управление проектом по разработке проектов машинного обучения. Анализ рисков в	(6 ч.) *Задание 2*. Оценка сроков и стоимости разработки проекта машинного обучения. Анализ рисков.	Изучение дополнительного материала по темам: Оценка сроков и стоимости разработки. Управление проектом по разработке проектов машинного обучения. Анализ рисков в проектах по машинному обучению. Постановка задач.

№, наименование темы	Содержание лекций (кол-во часов)	Наименование практических (семинарских занятий) (кол-во часов)	Виды СРС (кол-во часов)
	проектах по машинному обучению. Постановка задач. (2 ч.)	Формирование команды и бэклога проекта.	Формирование команды и бэклога проекта. (10 ч.)
Тема 3.3. Внедрение моделей машинного обучения	Шаблоны внедрения моделей машинного обучения. Разработка программного обеспечения с использованием моделей машинного обучения (2 ч.)	(6 ч.) *Задание 3*. Проектирование архитектуры системы машинного обучения. Внедрение моделей машинного обучения. Разработка, тестирование и развертывание системы машинного обучения.	Изучение дополнительного материала по темам: Шаблоны внедрения моделей машинного обучения. Разработка программного обеспечения с использованием моделей машинного обучения. Проектирование архитектуры системы машинного обучения. Внедрение моделей машинного обучения. Разработка, тестирование и развертывание системы машинного обучения. (10 ч.)

3. Условия реализации программы модуля

Организационно-педагогические условия реализации программы

Обучение по программе реализовано в формате смешанного обучения, с применением активных технологий совместного обучения в электронной среде (синхронные и асинхронные занятия). Лекционный материал представляется в виде синхронных лекций, текстовых материалов и презентаций. Данные материалы сопровождаются практическими заданиями. Изучение теоретического материала (СРС) предполагается до и после синхронной части работы.

Материально-технические условия реализации программы

Синхронные занятия реализуются на базе инструментов видеоконференцсвязи и включают в себя лекционные и практические занятия. Для проведения синхронных занятий (вебинаров) применяется программа видеоконференцсвязи. При проведении лекций, практических занятий, самостоятельной работы слушателей используется следующее оборудование: компьютер с наушниками или аудиоколонками, микрофоном и веб-камерой. Программное обеспечение (обновленное до последней версии): браузер Google Chrome, Anaconda, PyCharm Community.

Учебно-методическое и информационное обеспечение программы

Дисциплина может быть реализована как очно, так и заочно, в том числе, с применением дистанционных образовательных технологий. Она включает занятия лекционного типа, интерактивные формы обучения, семинарские, практические занятия.

Содержание комплекта учебно-методических материалов

По данной дисциплине имеется учебно-методический комплекс (УМК), который содержит: систему навигации по дисциплине (учебно-тематический план, график работы по дисциплине, сведения о результатах обучения), текстовые материалы к лекциям, практические задания, списки основной и дополнительной литературы.

Литература

1. Бурков А. Инженерия машинного обучения / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2022. – 306 с.

4. Оценка качества освоения программы модуля (формы аттестации, оценочные и методические материалы)

Форма аттестации по дисциплине — зачет. Зачет выставляется за выполненные практические задания.

Перечень заданий

Практические задания дисциплины

Практическая работа № 1. Проектирование системы интеллектуального анализа больших данных

Цель работы: приобретение навыков проектирования систем обработки данных, формализации постановки задач на разработку интеллектуальных систем, навыки анализа предметной области.

Общая последовательность действий

1. Определить постановку задачи в терминах предметной области и бизнес-метрики.
2. Формализовать постановку задачи и сформировать критерии и метрики оценки качества моделей машинного обучения для оценки результата.
3. Провести анализ аналогов.
4. Сформулировать требования к программной системе и данным для обучения моделей машинного обучения.
5. Спроектировать систему в любой нотации на выбор студента (диаграмма классов, функциональная декомпозиция, моделирование бизнес-процессов и другие).

Практическая работа № 2. Управление проектом машинного обучения

Цель: знакомство с инструментами организации процесса управления проектом по разработке систем машинного обучения, получение навыков формирования бэклога, оценки сроков и стоимости проекта.

Общая последовательность действий

1. Сформировать бэклог проекта и расставить приоритеты реализации.

2. Провести анализ рисков.
3. Оценить трудоемкость задач в соответствии с выбранной методологией (использовать Story Points / метод PERT / оценку по наихудшему-наилучшему случаю).
4. Оценить себестоимость каждого этапа и всего проекта.

Практическая работа № 3. Внедрение моделей машинного обучения

Цель работы: изучение инструментов для автоматизации оценки качества моделей машинного обучения, получение навыков внедрения моделей машинного обучения в программную систему.

Общая последовательность действий

1. Создать удаленный репозиторий для проекта.
2. Настроить систему версионирования данных и моделей DVC.
3. Провести серию экспериментов, используя как минимум три различные модели машинного обучения для решения поставленной задачи.
4. Выполнить перемещение по истории экспериментов с целью выбора конкретной модели.
5. Выполнить автоматическое тестирование качества работы модели машинного обучения
6. Разработать программное обеспечение в соответствии с задачей.
7. Интегрировать модель машинного обучения в приложение.

Задания для самостоятельной работы

В самостоятельные работы входит изучение дополнительного материала по темам дисциплины и закрепление заданий с практических занятий.

Критерии оценивания заданий и/или контрольных вопросов

Баллы	3 балла	4 балла	5 балла
Критерий	Задание выполнено частично, требует серьезной доработки	Задание выполнено, но требует некоторой доработки	Задание выполнено полностью, не требует доработки

Программу составили:

Старший преподаватель А.С. Михалев

Старший преподаватель П.В. Пересунько

Ассистент Е.О. Пересунько

Руководитель программы:

Д-р. техн. наук, профессор О.А. Антамошкин