

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Астраханский государственный университет имени В. Н. Татищева»
(Астраханский государственный университет им. В. Н. Татищева)

СОГЛАСОВАНО
Руководитель ОПОП

Р.Ю. Демина
«13» июня 2024 г.

УТВЕРЖДАЮ
Заведующий кафедрой ИТ

А.Н. Марьенков
«13» июня 2024 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Программирование для анализа данных

наименование

Составитель(-и)	Мартьянова А.Е., к.т.н., доцент кафедры
Направление подготовки	информационной безопасности
Направленность (профиль) ОПОП	09.03.02 Информационные системы и
Квалификация (степень)	технологии
Форма обучения	«Безопасность информационных систем»
Год приема	бакалавр
Курс	очно-заочная
Семестры	2021
	4
	8

Астрахань, 2024

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Целями освоения дисциплины (модуля) «Программирование для анализа данных» являются освоение и развитие навыков программирования на языке Python, представления о сборе, обработке и анализе данных, введение в автоматизированные методы работы с данными - машинное обучение и нейронные сети.

1.2. Задачи освоения дисциплины (модуля):

- освоение продвинутых методов исследования взаимосвязей между показателями, характеризующими объекты в социально-экономических исследованиях;
- освоение продвинутых методов распознавания образов и типологизации объектов;
- освоение продвинутых методов оптимизации представления информации об объектах;
- освоение современных пакетов прикладных программ, реализующих алгоритмы многомерного анализа данных;
- приобретение навыков содержательной интерпретации результатов исследования.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

2.1. Учебная дисциплина (модуль) Б1.В.Д.05.02 «Программирование для анализа данных» относится к части, формируемая участниками образовательных отношений и осваивается в 8 семестре, общая трудоемкость дисциплины — 3 ЗЕ, 108 часов, итоговая форма контроля — экзамен.

Дисциплина (модуль) «Программирование для анализа данных» встраивается в структуру ОПОП ВО (последовательность в учебном плане) как с точки зрения преемственности содержания, так и с точки зрения непрерывности процесса формирования компетенций выпускника. Обязательными знаниями для успешного освоения курса является основы математической статистики и теории вероятности, основ программирования на языке Python.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы следующие знания, умения и навыки, формируемые предшествующими учебными дисциплинами (модулями):

- Математические основы информационных технологий и вычислительной техники
- Основы программирования.

знания:

- роль и значение информационных ресурсов в современном обществе,
- виды и формы информации,
- современные информационные технологии обработки информации,
- этапы и методы ее обработки информации,

умения:

- применять компьютерную технику и информационные технологии для обработки информации, и решения практических задач,

навыки:

- владения инструментальных средств информационных технологий обработки информации,
- владения инфокоммуникационных технологий.

2.3. Последующие учебные дисциплины (модули) и (или) практики, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной (модулем):

- Машинное обучение.
- Интеллектуальные системы и технологии.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс освоения дисциплины (модуля) направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки (специальности):

в) профессиональной (ПК):

ПК-2. Способен разрабатывать программное обеспечение, включая проектирование, отладку, проверку работоспособности и модификацию ПО.

Таблица 1 - Декомпозиция результатов обучения

Код компетенции	Планируемые результаты обучения по дисциплине (модулю)		
	Знать (1)	Уметь (2)	Владеть (3)
ПК-2: способен разрабатывать программное обеспечение, включая проектирование, отладку, проверку работоспособности и модификацию ПО	ИПК 2.1. Знать: современные информационные технологии разработки, отладки, проверки работоспособности, модификации программного обеспечения	ИПК 2.2. Уметь: осуществлять выбор информационных технологий для решения задач по разработке, отладке, проверке работоспособности, модификации программного обеспечения	ИПК 2.3. Владеть: навыками разработки, отладки, проверки работоспособности, модификации программного обеспечения с использованием современных информационных технологий

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Объем дисциплины (модуля) составляет 3 зачетных единицы, 108 часов, в том числе 36 часов, выделенных на контактную работу обучающихся с преподавателем выделено (из них 18 часов – лекции, 18 часов – лабораторные работы, и 72 часа – на самостоятельную работу обучающихся).

Таблица 2 - Структура и содержание дисциплины (модуля)

№ п/п	Раздел, тема дисциплины (модуля)	Семестр	Неделя семестра	Контактная работа (в часах)			Самост. работа		Формы текущего контроля успеваемости, форма промежуточной аттестации (по семестрам)
				Л	ПЗ	ЛР	КР	СР	
1	Основы программирования на языке Python (refresher)	8	1-2	2		2		10	Лабораторная работа №1
2	Математический аппарат (refresher)	8	3-5	3		3		10	Лабораторная работа №2
3	Введение в модуль NumPy	8	6-8	3		3		10	Лабораторная работа №3
4	Основы	8	9-10	2		2		10	Лабораторная работа №4.

	работы с Pandas								Устный опрос
5	Разведывательный анализ данных	8	11-13	3		3		10	Лабораторная работа №5
6	Визуализация данных. Представление результатов	8	14-15	2		2		10	Лабораторная работа №6
7	Работа с текстовыми данными. Текстовый анализ	8	16-18	3		3		12	Лабораторная работа №7. Устный опрос
ИТОГО			108	18		18		72	Экзамен

Условные обозначения:

Л – лекция; ПЗ – практическое занятие, ЛР – лабораторная работа; КР – курсовая работа; СР – самостоятельная работа.

Таблица 3 - Матрица соотношения разделов, тем учебной дисциплины (модуля) и формируемых в них компетенций

Раздел, тема дисциплины (модуля)	Кол-во часов	Код компетенции	Общее количество компетенций
		ПК - 2	
Основы программирования на языке Python (refresher)	14	+	1
Математический аппарат (refresher)	16	+	1
Введение в модуль NumPy	16	+	1
Основы работы с Pandas	14	+	1
Разведывательный анализ данных	16	+	1
Визуализация данных. Представление результатов	14	+	1
Работа с текстовыми данными. Текстовый анализ	18	+	1
Итого:	108		

Краткое содержание каждой темы дисциплины (модуля)

Лекция 1. Основы программирования на языке Python (refresher).

Введение в анализ данных на языке Python. Почему Python становится стандартом для работы с большими данными. Прикладные задачи политологов, для решения которых подходит язык Python. Основы программирования на языке Python: типы данных и методы работы с ними (переменные, листы, словари, кортежи).

Лекция 2. Математический аппарат (refresher).

Математический аппарат для анализа данных: векторы, матрицы, функции и производные. Основы программирования на языке Python: циклы функции, знание синтаксиса языка.

Лекция 3. Введение в модуль NumPy.

Введение в модуль для работы с числовыми данными NumPy (Numerical Python) Особенности типов данных в NumPy. Работа с векторами и матрицами. Вычисление главных статистических метрик с помощью NumPy (среднее, медиана, мода, дисперсия).

Лекция 4. Основы работы с Pandas.

Введение в модуль для работы с табличным представлением данных Pandas. Преобразование словарей в табличный формат Pandas, загрузка данных из внешних источников. Особенности фильтрации и обращения к данным. Работа с табличными данными в Pandas на примере.

Лекция 5. Разведывательный анализ данных.

Регрессионные модели. Метод наименьших квадратов. Логистическая регрессия. Проведение разведывательного анализа данных в Pandas на примере.

Лекция 6. Визуализация данных. Представление результатов

Введение в визуализацию данных. Нюансы визуализации данных и принципы человеческого восприятия. Правила создания хороших визуализаций. Создание различных видов визуализаций на синтетических данных и тренировочных наборах данных.

Лекция 7. Работа с текстовыми данными. Текстовый анализ.

Введение в анализ текста. Применение в политологии. Особенности подготовки данных. Создание интерактивных визуализаций и отчетов с помощью инструмента Plotly.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРЕПОДАВАНИЮ И ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1. Указания для преподавателей по организации и проведению учебных занятий по дисциплине (модулю)

При подготовке к лекционным занятиям необходимо воспользоваться учебно-методической литературой (основной) из п.8.

При подготовке к лабораторным занятиям необходимо воспользоваться учебно-методической литературой (дополнительной) из п.8.

5.2. Указания для обучающихся по освоению дисциплины (модулю)

Во время самостоятельной работы необходимо воспользоваться учебно-методической литературой из п.8 (основной), (дополнительной), Интернет-ресурсами.

Таблица 4 - Содержание самостоятельной работы обучающихся

Номер раздела (темы)	Темы/вопросы, выносимые на самостоятельное изучение	Кол-во часов	Формы работы
Основы программирования на языке Python (refresher)	Подготовка отчета по лабораторной работе 1	10	Внеаудиторная, изучение учебных пособий
Математический аппарат (refresher)	Подготовка отчета по лабораторной работе 2	10	Внеаудиторная, изучение учебных пособий
Введение в модуль NumPy	Подготовка отчета по лабораторной работе 3	10	Внеаудиторная, изучение учебных пособий
Основы работы с Pandas	Подготовка отчета по лабораторной работе 4. Подготовка к устному опросу	10	Внеаудиторная, изучение учебных пособий
Разведывательный	Подготовка отчета по лабораторной	10	Внеаудиторная,

Номер раздела (темы)	Темы/вопросы, выносимые на самостоятельное изучение	Кол-во часов	Формы работы
анализ данных	работе 5		изучение учебных пособий
Визуализация данных. Представление результатов	Подготовка отчета по лабораторной работе 6	10	Внеаудиторная, изучение учебных пособий
Работа с текстовыми данными. Текстовый анализ	Подготовка отчета по лабораторной работе 7. Подготовка к устному опросу	12	Внеаудиторная, изучение учебных пособий

5.3. Виды и формы письменных работ, предусмотренных при освоении дисциплины, выполняемые обучающимися самостоятельно – подготовка реферата.

Не предусмотрено.

6. ОБРАЗОВАТЕЛЬНЫЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

При реализации различных видов учебной работы по дисциплине могут использоваться электронное обучение и дистанционные образовательные технологии.

6.1. Образовательные технологии

Учебные занятия по дисциплине могут проводиться с применением информационно-телекоммуникационных сетей при опосредованном (на расстоянии) интерактивном взаимодействии обучающихся и преподавателя в режимах on-line в формах: видеолекций, лекций-презентаций, видеоконференции, собеседования в режиме чат, форума, чата, выполнения виртуальных практических и/или лабораторных работ и др.

Максимальный объем занятий обучающегося с применением электронных образовательных технологий не должен превышать 25%.

Таблица 5 – Образовательные технологии, используемые при реализации учебных занятий

Раздел, тема дисциплины (модуля)	Форма учебного занятия		
	Лекция	Практическое занятие, семинар	Лабораторная работа
Основы программирования на языке Python (refresher)	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Математический аппарат (refresher)	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Введение в модуль NumPy	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Основы работы с Pandas	Лекция - презентация	Не предусмотрено	Выполнение лабораторной

			работы
Разведывательный анализ данных	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Визуализация данных. Представление результатов	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы
Работа с текстовыми данными. Текстовый анализ	Лекция - презентация	Не предусмотрено	Выполнение лабораторной работы

Учебные занятия по дисциплине могут проводиться с применением информационно-телекоммуникационных сетей при опосредованном (на расстоянии) интерактивном взаимодействии обучающихся и преподавателя в режимах on-line в формах: видеолекций, лекций-презентаций, видеоконференции, собеседования в режиме чат, форума, чата, выполнения виртуальных практических и/или лабораторных работ и др.

Максимальный объем занятий обучающегося с применением электронных образовательных технологий не должен превышать 25%.

6.2. Информационные технологии

- использование возможностей интернета в учебном процессе (использование сайта преподавателя (рассылка заданий, предоставление выполненных работ, ответы на вопросы, ознакомление обучающихся с оценками и т.д.));
- использование электронных учебников и различных сайтов (например, электронных библиотек, журналов и т. д.) как источников информации;
- использование возможностей электронной почты преподавателя;
- использование средств представления учебной информации (электронных учебных пособий и практикумов, применение новых технологий для проведения очных (традиционных) лекций и семинаров с использованием презентаций и т. д.);
- использование интегрированных образовательных сред, где главной составляющей являются не только применяемые технологии, но и содержательная часть, т. е. информационные ресурсы (доступ к мировым информационным ресурсам, на базе которых строится учебный процесс);
- использование виртуальной обучающей среды (LMS Moodle «Электронное обучение») или иных информационных систем, сервисов и мессенджеров.

6.3. Программное обеспечение, современные профессиональные базы данных и информационные справочные системы

6.3.1. Программное обеспечение

Наименование программного обеспечения	Назначение
AdobeReader	Программа для просмотра электронных документов
Платформа дистанционного обучения LMS Moodle	Виртуальная обучающая среда

Mozilla FireFox	Браузер
Microsoft Office 2013, Microsoft Office Project 2013 , Microsoft Office Visio 2013	Офисная программа
7-zip	Архиватор
Microsoft Windows 7 Professional	Операционная система
Kaspersky Endpoint Security	Средство антивирусной защиты

Среда разработки Anaconda3 (64 –bit) (в свободном доступе)

6.3.2. Современные профессиональные базы данных и информационные справочные системы

1. Электронный каталог Научной библиотеки АГУ на базе MARKSQL НПО «Информ-систем»: <https://library.asu.edu.ru>.
2. Электронный каталог «Научные журналы АГУ»: <http://journal.asu.edu.ru/>.
3. Универсальная справочно-информационная полнотекстовая база данных периодических изданий ООО «ИВИС»: <http://dlib.eastview.com/>.
4. Электронно-библиотечная система elibrary. <http://elibrary.ru>
5. Справочная правовая система КонсультантПлюс: <http://www.consultant.ru>
6. Информационно-правовое обеспечение «Система ГАРАНТ»: <http://garant-astrakhan.ru>

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

7.1. Паспорт фонда оценочных средств

При проведении текущего контроля и промежуточной аттестации по дисциплине (модулю) «Программирование для анализа данных» проверяется сформированность у обучающихся компетенций, указанных в разделе 3 настоящей программы. Этапность формирования данных компетенций в процессе освоения образовательной программы определяется последовательным освоением дисциплин (модулей) и прохождением практик, а в процессе освоения дисциплины (модуля) – последовательным достижением результатов освоения содержательно связанных между собой разделов, тем.

Таблица 6 - Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств

№ п/п	Контролируемые разделы, темы дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
1	Основы программирования на языке Python (refresher).	ПК 2	Лабораторная работа №1
2	Математический аппарат (refresher).	ПК 2	Лабораторная работа №2
3	Введение в модуль NumPy.	ПК 2	Лабораторная работа №3
4	Основы работы с Pandas.	ПК 2	Лабораторная работа №4. Вопросы для обсуждения.
5	Разведывательный анализ данных.	ПК 2	Лабораторная работа №5
6	Визуализация данных.	ПК 2	Лабораторная работа

	Представление результатов		№6
7	Работа с текстовыми данными. Текстовый анализ.	ПК 2	Лабораторная работа №7. Вопросы для обсуждения

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

Таблица 7 - Показатели оценивания результатов обучения в виде знаний

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует глубокое знание теоретического материала, умение обоснованно излагать свои мысли по обсуждаемым вопросам, способность полно, правильно и аргументированно отвечать на вопросы, приводить примеры
4 «хорошо»	демонстрирует знание теоретического материала, его последовательное изложение, способность приводить примеры, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует неполное, фрагментарное знание теоретического материала, требующее наводящих вопросов преподавателя, допускает существенные ошибки в его изложении, затрудняется в приведении примеров и формулировке выводов
2 «неудовлетворительно»	демонстрирует существенные пробелы в знании теоретического материала, не способен его изложить и ответить на наводящие вопросы преподавателя, не может привести примеры

Таблица 8 - Показатели оценивания результатов обучения в виде умений и владений

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы
4 «хорошо»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует отдельные, несистематизированные навыки, не способен применить знание теоретического материала при выполнении заданий, испытывает затруднения и допускает ошибки при выполнении заданий, выполняет задание при подсказке преподавателя, затрудняется в формулировке выводов
2 «неудовлетворительно»	не способен правильно выполнить задание

7.3. Контрольные задания и иные материалы, необходимые для оценки результатов обучения по дисциплине (модулю)

Основы программирования на языке Python (refresher).

Лабораторная работа 1. Основы программирования на языке Python: типы данных и методы работы с ними (переменные, листы, словари, кортежи).

Цель: ознакомиться с основами языка Python, получить умения для выполнения дальнейших лабораторных работ.

Задачи:

- изучить типизацию данных;
- рассмотреть «ветвление» в Python;
- отработать задачи с использованием конструкции «try-except»;

В ней обучающиеся поэтапно проходят все основные аспекта языка, такие как:

1. Типизация данных.
2. Ветвления.
3. Исключения.
4. Переменные, листы, словари, кортежи

Математический аппарат (refresher).

Лабораторная работа 2. Задание 1. Основы программирования на языке Python: циклы функции, знание синтаксиса языка.

Цель: ознакомиться с основами языка Python: циклы функции, знание синтаксиса языка.

Задачи:

- научиться пользоваться циклами «for» и «while»;
- разобрать функции и пространства имён.

В ней обучающиеся поэтапно проходят все основные аспекта языка, такие как:

1. Пространство имён.
2. Функции.
3. Циклы.

Задание 2. Классификация данных

Цель: научиться работать с данными при помощи визуальных инструментов и разобрать азы классификации при помощи построения простейшего классификатора со статичными параметрами.

Задачи:

- научиться анализировать данные;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- построить классификатор на основе данных полученных при анализе;
- научиться калибровать нейросеть для получения более точных ответов.

В этой лабораторной работе начинается изучение принципов машинного обучения. В ней обучающиеся будут учиться работать с графической информацией и производить анализ полученных данных и самостоятельно строить простейший классификатор. Данная работа поможет понять простейшие понятия и принципы работы более сложных классификаторов.

Введение в модуль NumPy.

Лабораторная работа 3. Задание 1. Особенности типов данных в NumPy. Работа с векторами и матрицами. Вычисление главных статистических метрик с помощью NumPy (среднее, медиана, мода, дисперсия).

Цель: ознакомиться с основами NumPy, работать с векторами и матрицами, статистическими метриками.

Задачи:

- изучить особенности типов данных в NumPy,
- научиться работать с векторами и матрицами.

- получить умения вычислять главные статистические метрики с помощью NumPy (среднее, медиана, мода, дисперсия).

Задание 2. Классификация методом "К-ближайших соседей"

Цель: изучить метод простейший метод классификации данных «К-ближайших соседей» и научиться производить оценку данных с помощью визуальных инструментов Python.

Задачи:

- детально разобрать метод машинного обучения «К-ближайших соседей»;
- научиться работать с информацией;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- научиться калибровать нейросеть для получения более точных ответов.

В этой лабораторной работе будет рассматриваться метод машинного обучения «К-ближайших соседей», а также его положительные и отрицательные стороны; рассмотрим библиотеку для визуализации данных и математического анализа «matplotlib»; будем производить калибровку нейронной сети для получения максимального процента правильных ответов, а также обсудим проблему переобучения в нейросетях.

Основы работы с Pandas.

Лабораторная работа № 4. Работа с табличными данными в Pandas на примере.

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для визуализации данных в датасетах.

Задачи:

- получить умения по использованию библиотеки Pandas;
- сформировать понятия о DataFrame и Series;
- научиться строить графики с помощью scatter matrix (матрица рассеивания) и matplotlib.

Pandas — это библиотека языка Python, которая позволяет создавать объекты, в которых данные хранятся в табличной форме. Это основной структурный объект необходимый для анализа и очистки данных перед их использованием в обучении модели. Это первая лабораторная работа, связанная с Pandas. В ней обучающиеся разберут структуру и основные функции данной библиотеки.

Вопросы для обсуждения

- 1) Введение в программирование на языке Python.
- 2) Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, Pandas.

Разведывательный анализ данных.

Лабораторная работа № 5. Проведение разведывательного анализа данных в Pandas на примере.

Задание 1. Анализ данных с помощью Pandas»

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для анализа данных в датасетах.

Задачи:

- получить умения по использованию библиотеки Pandas;
- научиться анализировать и обрабатывать данные с помощью Pandas;
- закрепить умения визуализации в Pandas.

Обучающиеся переходят к анализу данных с помощью библиотеки Pandas. В данной работе обучающиеся научатся делать срезы, группировки, индексацию данных по разным параметрам, а также строить графики по созданным срезам. Знания и умения, полученные в ходе выполнения работы помогут понять, некоторые механизмы машинного обучения.

Задание 2. Линейная регрессия

Цель: понять и научиться применять метод линейной регрессии в машинном обучении.

Задачи:

- изучить модель линейного регрессора;
- произвести обучение модели;
- рассмотреть особенности данного метода машинного обучения;
- произвести предсказание на основе созданной модели.

Линейная регрессия является одним из самых универсальных методов машинного обучения, т.к. может принимать на вход данные различных типов и структур. Также данная модель достаточно устойчива к переобучению и имеет большую скорость обработки данных. В этой лабораторной будут рассматриваться все нюансы и принципы работы этого метода классификации.

Визуализация данных. Представление результатов

Лабораторная работа № 6. Создание различных видов визуализаций на синтетических данных и тренировочных наборах данных.

Цель: научить обучающихся основам работы с машинным зрением и показать основные алгоритмы работы с ним.

Задачи:

- разобрать импорт и просмотр изображения;
 - разобрать кадрирование;
 - научиться изменять размер изображения;
 - научиться переворачивать изображение;
- рассмотреть способ преобразование изображения в черно-белое;
- научиться работать со сглаживанием и размытием;
 - изучить метод распознавания лиц.

В последние годы машинное зрение получило большую огласку и вызвало интерес со стороны не только ученых, но и различных инженеров и разработчиков «интеллектуальных» приложений. В лабораторной работе обучающиеся рассмотрят вариант библиотеки OpenCV, написанной на языке «C++» для Python, опробуем некоторый её функционал и протестируем классификатор для распознавания лиц.

Работа с текстовыми данными. Текстовый анализ.

Лабораторная работа № 7. Задание 1. Особенности подготовки данных. Создание интерактивных визуализаций и отчетов с помощью инструмента Plotly.

Цель: научиться работать с текстовыми данными.

Задачи:

- рассмотреть особенности подготовки данных,
- научиться создавать интерактивные визуализации и отчеты с помощью инструмента Plotly

Задание 2. Деревья решений

Цель: познакомить обучающихся с методом машинного обучения, построенном на деревьях решений, а также научить строить сами деревья.

Задачи:

- рассмотреть понятие дерева решений;
- рассмотреть варианты применения данной классификации;
- обучить модель на основе классов;
- отобразить дополнительный класс на модели и посмотреть результат;
- рассмотреть плюсы и минусы данной модели.

Одним из самых простых для понимания является классификатор «Дерево решений». Эта модель строит «дерево» классов, у каждого из которых есть свои параметры. Попадающие под эти параметры данные причисляются к тому или иному классу. Модель очень чувствительна к переобучению и данным, которые получает на вход. Лабораторная работа поможет нам разобраться во всех тонкостях данного метода и покажет все его нюансы.

Задание 3. Метод случайного леса

Цель: сформировать понятие случайного леса, а также научить обучающихся использовать данную модель для решения задач.

Задачи:

- рассмотреть понятие случайного леса;
- рассмотреть пример кода для решения простых задач;
- научить подбирать параметры модели для улучшения качества прогнозов модели.

Случайный лес является модифицированным вариантом «Древа решений» и не имеет всех его отрицательных особенностей, например, он не чувствителен к переобучению. В лабораторной работе будет рассмотрено строение данного метода классификации и его устройство.

Вопросы для обсуждения

- 1) Введение в машинное обучение. Модуль sklearn.
- 2) Задачи классификации и линейные модели. Деревья решений. Случайный лес. Ансамбли моделей. Введение в нейронные сети.

Перечень вопросов и заданий, выносимых на экзамен / зачёт / дифференцированный зачёт

Таблица 9 – Примеры оценочных средств с ключами правильных ответов

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
ПК-2: способен разрабатывать программное обеспечение, включая проектирование, отладку, проверку работоспособности и модификацию ПО				
1.	Задание закрытого типа	Что такое нормализация данных? 1. Усреднение данных 2. Преобразование категориальных признаков в численные 3. Преобразование численных признаков в категориальные 4. Подгонка под единую шкалу	3	5
2.		Укажите соответствие между типами входных/целевых признаков и диаграммой, которую целесообразно использовать для визуализации 1. Входной признак- категориальный, целевая переменная- категориальная 2. Входной признак- категориальный, Целевая переменная- числовая 3. Входной признак-	1-d 2-c 3-b 4-a	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
		числовой, Целевая переменная- категориальная 4. Входной признак- числовой, Целевая переменная- числовая a. Диаграмма рассеяния b. Диаграмма размаха c. График плотности d. Мозаичная диаграмма		
3.		Для разработки нейросетевых моделей используются библиотеки: 1. Pymorphy2 2. Keras 3. PyTorch 4. OpenCV	2,3	2
4.		Для реализации Web-интерфейсов приложений ИИ используются: 1. Flask 2. Git 3. Dash 4. Numba	1,3	3
5.		Какой алгоритм основан на гипотезе «Набор слабых обучающих алгоритмов способен создать сильный обучающий алгоритм»? 1. Бустинг 2. Случайный лес 3. Нейронные сети 4. Наивный Байес	1	3
6.	Задание открытого типа	Предположим, в обучающем множестве у некоторых объектов отсутствуют значения признаков. Какие варианты возможны в данной ситуации?	Можно удалить объекты с пропущенными значениями признаков. Можно преобразовать отсутствующие значения в значимые числа и специально созданную категорию. Можно заполнить значения признаков значением предшествующего экземпляра или медианой	5

№ п/ п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			столбца.	
7.		Что такое верность (ассурасу) классификации?	Верность-это доля правильно распознанных экземпляров.	5
8.		Какие признаки называются категориальными?	Признаки называются категориальными, если их можно отнести к какой-либо группе, но при этом не важен порядок	5
9.		Какая информация хранится в матрицах сопряженности по результатам тестирования классификатора?	Информация о правильно и неправильно распознанных объектах каждого класса: FP, TP, FN, TN	5
10.		Перечислите основные типы алгоритмов кластеризации	• Иерархический • k-средних • с-средних • Выделение связанных компонент • Минимальное покрывающее дерево • Послойная кластеризация	5

Вопросы к экзамену:

1. Язык Python и особенности его стиля программирования. Интерактивный режим Python. Ipython. Jupyter Notebook.
2. Синтаксис и управляющие конструкции языка Python. Переменные, значения и их типы. Типы данных в Python.
3. Встроенные операции и функции. Основные алгоритмические конструкции.
4. Условный оператор. Множественное ветвление.
5. Циклы и счетчики.
6. Определение функций. Параметры и аргументы. Вызовы функций. Оператор возврата. Конструкции *args, **kwargs.
7. Списки, кортежи и словари.
8. Операторы общие для всех типов последовательностей.
9. Специальные операторы и функции для работы со списками. Срезы.
10. Работа со словарями. Методы словарей.
11. Случайные числа. random, randrange, choice.
12. Функции обработки строк. join, replace, split.
13. Стандартная библиотека и pip. Модули и пакеты в Python. Основные стандартные модули.
14. Импортирование модулей. Создание собственных модулей и их импортирование. Специализированные модули и приложения.
15. Файлы и исключения. Работа с внешними источниками данных.
16. Исключения, обработка исключений, вызов исключений (try-except-finally).
17. Утверждения (assert). Открытие, чтение, запись. (open, инструкция with).
18. Работа с текстовыми файлами, xml и csv - файлами.
19. Функциональное программирование. Лямбда-функции.
20. Использование функций map, filter, reduce, zip.
21. Генераторы, декораторы, рекурсия.

22. Модификация функций с помощью декораторов.
23. Итерируемые объекты. Использование генераторов (yield).
24. ООП в Python. Классы, объекты и экземпляры классов. Наследование.
25. Магические методы. Переопределение операторов. Методы классов.
26. Инкапсуляция. Условно частные и строго частные методы.
27. Регулярные выражения. Использование регулярных выражений. Пакет re.
28. Наука о данных и Python. Библиотеки: NumPy, pandas, matplotlib, SciPy.
29. Основы NumPy: массивы и векторные вычисления.
30. Инструменты визуализации данных для Python.
31. Введение в API библиотеки matplotlib.
32. Библиотека pandas. Введение в структуры данных pandas.
33. Объекты Dataframe и Series.
34. Визуализация данных в pandas. Seaborn.
35. Агрегирование данных и групповые операции.
36. Сбор и подготовка данных в Python: извлечение данных с web-страниц (web-scraping). Библиотека BeautifulSoup.
37. Работа с динамическими сайтами с помощью Selenium.
38. Массовый скрепинг с помощью scrapy.
39. Работа со структурированными данными: JSON и XML.
40. Открытые API. Telegram API, VK API.

7.4. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине (модулю)

Таблица 10 – Технологическая карта рейтинговых баллов по дисциплине (модулю)

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Основной блок				
1.	<i>Выполнение лабораторной работы</i>	6/7	42	По расписанию
2.	<i>Ответ во время опроса</i>	4/2	8	
Всего			50	-
Блок бонусов				
3.	<i>Посещение занятий</i>	0,25/24	6	
4.	<i>Своевременное выполнение всех заданий</i>	0,5/8	4	
Всего			10	-
Дополнительный блок**				
5.	<i>Экзамен</i>		40	
Всего			40	-
ИТОГО			100	-

Таблица 11 – Система штрафов (для одного занятия)

Показатель	Балл
<i>Опоздание на занятие</i>	-0,5
<i>Нарушение учебной дисциплины</i>	-5
<i>Неготовность к занятию</i>	-1
<i>Пропуск занятия без уважительной причины</i>	-2

Таблица 12 – Шкала перевода рейтинговых баллов в итоговую оценку за семестр по дисциплине (модулю)

Сумма баллов	Оценка по 4-балльной шкале
90–100	5 (отлично)
85–89	4 (хорошо)
75–84	
70–74	
65–69	
60–64	3 (удовлетворительно)
Ниже 60	2 (неудовлетворительно)

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

8.1. Основная литература

1. Кузьмич, Р. И. Модификации метода логического анализа данных для задач классификации : монография / Р. И. Кузьмич, И. С. Масич. — Красноярск : Сибирский федеральный университет, 2018. — 181 с. — ISBN 978-5-7638-3698-1. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/84252.html>

2. Воронова, Л. И. Machine Learning: регрессионные методы интеллектуального анализа данных : учебное пособие / Л. И. Воронова, В. И. Воронов. — Москва : Московский технический университет связи и информатики, 2018. — 82 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/81325.html>

3. Ракитский, А. А. Методы машинного обучения : учебно-методическое пособие / А. А. Ракитский. — Новосибирск : Сибирский государственный университет телекоммуникаций и информатики, 2018. — 32 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/90591>.

4. Рашка С., Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / Рашка С. - М. : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0 - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785970604090.html>

8.2. Дополнительная литература

1. Федин, Ф. О. Анализ данных. Часть 1. Подготовка данных к анализу : учебное пособие / Ф. О. Федин, Ф. Ф. Федин. — Москва : Московский городской педагогический университет, 2012. — 204 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/26444.html>

2. Билл, Фрэнкс Укрощение больших данных : как извлекать знания из массивов информации с помощью глубокой аналитики / Фрэнкс Билл ; перевод А. Баранов. — Москва : Манн, Иванов и Фербер, 2014. — 340 с. — ISBN 978-5-00057-146-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/39433.html>

3. Флах П., Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Флах П. - М. : ДМК Пресс, 2015. - 400 с. - ISBN 978-5-97060-

273-7 - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785970602737.html>

8.3. Интернет-ресурсы, необходимые для освоения дисциплины (модуля)

1. Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента». Многопрофильный образовательный ресурс «Консультант студента» является электронной библиотечной системой, предоставляющей доступ через сеть Интернет к учебной литературе и дополнительным материалам, приобретенным на основании прямых договоров с правообладателями. Каталог в настоящее время содержит около 15000 наименований. www.studentlibrary.ru.

2. Kaggle. Система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению. www.kaggle.com

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Для проведения лекционных занятий необходима мультимедийная аудитория, оснащенная компьютерной презентационной техникой.

Для проведения лабораторных занятий необходима аудитория, оснащенная компьютерами.

Рабочая программа дисциплины (модуля) при необходимости может быть адаптирована для обучения (в том числе с применением дистанционных образовательных технологий) лиц с ограниченными возможностями здоровья, инвалидов. Для этого требуется заявление обучающихся, являющихся лицами с ограниченными возможностями здоровья, инвалидами, или их законных представителей и рекомендации психолого-медико-педагогической комиссии. Для инвалидов содержание рабочей программы дисциплины (модуля) может определяться также в соответствии с индивидуальной программой реабилитации инвалида (при наличии).