Приложение 6 ОП ВО

Федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Северо-Западный институт управления – филиал РАНХиГС

Кафедра бизнес-информатики

УТВЕРЖДЕНО

Директор СЗИУ РАНХиГС
А.Д. Хлутков

ПРОГРАММА МАГИСТРАТУРЫ
Аналитическое обеспечение информационной безопасности
(наименование образовательной программы)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ, реализуемой без применения электронного (онлайн) курса

Б1.В.09 Интеллектуальный анализ текстов и изображений
(код и наименование РПД)

38.04.05 Бизнес-информатика
(код,наименование направления подготовки)

очная
(форма обучения)

Год набора – 2024

Санкт-Петербург, 2024 г.

Автор–составитель:
Кандидат технических наук, доцент кафедры бизнес-информатики Буров Сергей Александрович

Заведующий кафедрой «Бизнес-информатика»
Двн, профессор Наумов Владимир Николаевич

РПД «Интеллектуальный анализ данных, текстов, изображений» одобрена протоколом заседания кафедры бизнес-информатики № 6 от 06.03.2023 г.

СОДЕРЖАНИЕ

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения программы

Дисциплина «Интеллектуальный анализ данных, текстов и изображений» обеспечивает овладение следующими компетенциями:

Таблица 1

Код компетенции	Наименование Компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенции
ПКС-2	Способен обосновывать подходы, используемые в бизнес-анализе, руководить и управлять бизнес-анализом с использованием информационно-коммуникационных технологий	ПКС-2.1	Способен использовать современные методы, информационные технологии, программный инструментарий в объеме, необходимом для решения задач бизнес-аналитики, использовать англоязычную документацию и справочные системы

В результате освоения дисциплины у магистрантов должны быть сформированы компетенции:

Таблица 2

ОТФ/ТФ (при наличии профстандарта)/ профессиональные действия	Код этапа освоения компетенции	Результаты обучения
Управление бизнес-анализом	ПКС-2.1	на уровне знаний: **Знать:** – методы анализа данных и машинного обучения; – возможности программных средств статистической обработки и интеллектуального анализа данных
		на уровне умения: **Уметь:** – применять программные средства анализа данных, поддержки принятия решений; **Владеть:** современными инфокоммуникационными технологиями;

2.Объем и место дисциплины в структуре ОП ВО

Объем дисциплины

Общая трудоемкость дисциплины составляет 2 зачетные единицы /72 часа.

Таблица 3

Вид работы	Трудоемкость (акад/астр.часы)
Общая трудоемкость	72/54
Контактная работа с преподавателем	32/24
Лекции	16/12
Практические занятия	16/12

Лабораторные занятия	
Самостоятельная работа	74
Консультация	2/1,5
Формы текущего контроля	УО, Зад, Т
Форма промежуточной аттестации	Зачет

Место дисциплины в структуре ОП ВО

Дисциплина «Интеллектуальный анализ данных, текстов и изображений» относится к части, формируемой участниками образовательных отношений. Преподавание дисциплины опирается на знания, полученные в ходе изучения дисциплины Б1.В.02 «Математические методы статистической обработки и анализа данных», ФТД.02 «Предсказательная аналитика», изучаемые в первом семестре. Дисциплина изучается во втором семестре первого года обучения.

В свою очередь она создаёт необходимые предпосылки для освоения программ таких дисциплин, как Б1.О.07 «Аналитическая поддержка принятия решений».

Дисциплина закладывает теоретический и методологический фундамент для овладения умениям и навыками в ходе Б2.О.01(У) «Проектно-аналитическая практика» и Б2.О.02 (Н) «Научно-исследовательская работа».

Знания, умения и навыки, полученные при изучении дисциплины, используются студентами при выполнении выпускных квалификационных работ.

3.Содержание и структура дисциплины

3.1.Структура дисциплины

Очная форма обучения

№ п/п	Наименование тем	Объем дисциплины, час.						Форма текущего контроля успеваемости**, промежуточной аттестации***
		Всего	Контактная работа обучающихся с преподавателем по видам учебных занятий			СР		
			Л	ПЗ	КСР	СРО	СП	
Тема 1	Введение в интеллектуальный анализ данных. Основы интеллектуального анализа данных на Python.	16	4	4		8		УО/Зад/Т
Тема 2	Основы машинного обучения и нейронных сетей	18	4	4		10		УО/Зад/Т
Тема 3	Основы интеллектуального анализа изображений	18	4	4		10		УО/Зад
Тема 4	Основы интеллектуального анализа текстов	20	4	4		12		УО/Зад
Промежуточная аттестация					2*			Зачет
Всего (акад./астр. часы):		72/54	16/12	16/12	2/1,5	40/30		

Примечание:

2* - консультация, не входящая в общий объем дисциплины

Используемые сокращения:

Л – занятия лекционного типа (лекции и иные учебные занятия, предусматривающие

преимущественную передачу учебной информации педагогическими работниками организации и (или) лицами, привлекаемыми организацией к реализации образовательных программ на иных условиях, обучающимся) ;
ПЗ – практические занятия (виды занятия семинарского типа за исключением лабораторных работ) ;
КСР – индивидуальная работа обучающихся с педагогическими работниками организации и (или) лицами, привлекаемыми организацией к реализации образовательных программ на иных условиях (в том числе индивидуальные консультации) ;
СР – самостоятельная работа, осуществляемая без участия педагогических работников организации и (или) лиц, привлекаемых организацией к реализации образовательных программ на иных условиях;
СП – самопроверка;
СРО – самостоятельная работа обучающегося
контрольные работы (К), опрос (О), тестирование (Т)

3.2. Содержание дисциплины

Тема 1. Введение в интеллектуальный анализ данных. Основы интеллектуального анализа данных на Python

Введение. Содержание интеллектуального анализа данных. Среда разработки. Платформа Anaconda. Общая характеристика языка Python. Сравнительный анализ Phyton, R. Среда разработки. Платформа Anaconda. Основы синтаксиса языка. Переменные, ключевые слова. Основы программирования на языке Python. Типы данных. Функциональность для работы с данными. Установка пакетов научных вычислений на Python. Тензоры нулевого, первого, второго, третьего и высших порядков. Операции над тензорами и манипулирование тензорами с помощью NumPy. Интеллектуальный анализ данных с использованием библиотеки Pandas.

Тема 2. Основы машинного обучения и нейронных сетей

Машинное обучение в задачах интеллектуального анализа данных, текстов и изображений. Решение задач классификации и кластеризации для интеллектуального анализа данных. Деревья решений и ансамблевые методы машинного обучения. Библиотеки Keras, TensorFlow, Pytorch. Нейронные сети в задачах интеллектуального анализа данных. Представление данных для нейронных сетей. Понятие градиента. Стохастический градиент. Понятие поверхностного и глубокого обучения. Свёточные нейронные сети. Операция свертывания. Шаблоны свертки. Операция max pooling, padding. Архитектура сверточных сетей. Обучение и проверка сверточных нейронных сетей. Разработка нейронных сетей с использованием TensorFlow, Keras и PyTorch.

Тема 3. Основы интеллектуальный анализа изображений

Методы переноса знаний, локальные дескрипторы и структурные методы для решения задач распознавания изображений. Методы обработки и фильтрации изображений. Библиотека OpenCV. Решение задача классификации и кластеризации при интеллектуальном анализе изображений. Детектирование и сегментация объектов. Синтез изображений. Глубокое обучение в технологиях компьютерного зрения. Решение задачи распознавания рукописного текста на основе набора данных Mnist. Обучение и проверка сверточных нейронных сетей.

Тема 4. Основы интеллектуального анализа текстов

Основные этапы и методы интеллектуального анализа текстов. Классификация в текстовом анализе. Кластеризация для выявления сходств в тексте. Решение задачи

распознавания рукописного текста на основе набора данных MNIST. Глубокое обучение для текста и последовательностей. Работа с текстами. Прямое кодирование слов и символов. Токенизация. Латентно-семантический анализ. TF-IDF. N-граммы. Мешок слов. Векторное представление слов. Слой embedding. Рекуррентные нейронные сети в задачах интеллектуального анализа текстов. Слои LSTM, GRU.

4.Материалы текущего контроля успеваемости обучающихся и фонд оценочных средств промежуточной аттестации по дисциплине

4.1. Формы и методы текущего контроля успеваемости обучающихся и промежуточной аттестации.

В ходе реализации дисциплины используются следующие методы текущего контроля успеваемости обучающихся:

Таблица 4.1

Тема (раздел)	Формы текущего контроля успеваемости
Тема 1. Введение в интеллектуальный анализ данных. Основы интеллектуального анализа данных на Python.	УО/Зад/Т
Тема 2. Основы машинного обучения и нейронных сетей	УО/Зад
Тема 3. Основы интеллектуального анализа изображений	УО/Зад
Тема 4. Основы интеллектуального анализа текстов	УО/Зад

4.1.2 Зачет проводится по итогам текущего контроля

При выставлении зачета учитывается активность магистрантов в течение занятий, выполнение задач, решаемых в ходе занятий.

4. 2. Типовые материалы текущего контроля успеваемости обучающихся.

Типовые оценочные материалы по теме 1

Задания по теме 1

Задание1.

Построить диаграммы функций в разных стилях согласно блокноту Task_2.ipynb.

Задание 2. Модуль numpy

Сгенерировать последовательность равномерно распределенных случайных чисел, в диапазоне от 0 до 10. Размер последовательности 100. Аналогично сгенерировать еще одну последовательность целых чисел, в диапазоне от 0 до 20 также состоящую из 100 чисел. Используя функцию linspace() сгенерировать последовательность из 100 числе, заключенную от 0 до 10. Используя оператор цикла for и итератор i сформировать последовательность из 100 чисел от -1 до 9. Построить точечные, линейные диаграммы для полученных последовательностей.

Построить матричную диаграмму, используя стиль matlab. Задавать панели с помощью subplot. Применить средства управления цветом, типом линии, маркерами.

Задание 3. Модуль pandas

Загрузить набор данных. Файл данных хранится в moodle.

```
sp500=pd.read_csv(filepath_or_buffer="спецификация файла",sep=',',
            usecols=['Symbol','Sector','Price','Book Value'],
            index_col='Symbol')
sp500.head()
```

Провести анализ файла.

С помощью метода iloc отобразить только 10,20,30,40,50, 60 и 70 элементы

Отобразить только первые два столбца набора данных

С помощью метода sample выбрать 20 кампаний.

Задание 4.

Исследуйте возможности библиотеки Pandas самостоятельно согласно блокноту Task_vk.ipynb. В данном задании предлагается проанализировать базу данных пользователей социальной сети «Вконтакте». База данных загружается из файла «vk_main.csv». Данный файл должен находится в одном каталоге с этим блокнотом, или по пути, указанному пр загрузке «dataframe».

Вычислите среднее количество друзей «friends_cnt» у людей с маркетинга «is_bmm», у которых не указан инстаграм «instagram_dummy».

Вычислите количество юношей, которые знают английский «english_dummy» или у которых страница закрыта «is_closed» (также найти долю таких парней от общего числа парней).

Обработайте пропуски в столбцах «city» и «followers_cnt» (вывести датафрейм, не содержащий пропусков в этих столбцах).

Вычислите количество пропусков в «followers_cnt» и заполните их медианой.

В столбцах «photos_cnt», «videos_cnt» и «wall_post_cnt» вывести количество(без NaN), среднее, стандартное отклонение, минимум, максимум за одну команду.

Задание 5. Создание функций и классов

Создайте функцию, которая решает систему алгебраических уравнений методов обратной функции, если определитель матрицы не равен нулю. В противном случае выдается сообщение о том, что система уравнений не определена. Проверить решение на нескольких вариантах исходных данных. Сравнить с результатами решения, полученными вручную.

Постройте иерархию классов «Автомобиль». В иерархию классов включить классы: грузовые автомобили; легковые автомобили, гоночные автомобили, транспорт. В класс легковых автомобилей включить классы: седан, лимузин, купе, хетчбек, универсал, фургон, микроавтобус. Задать атрибуты «модель», «год выпуска». Задать свойства и методы для седана. Создать экземпляры седана

Типовые вопросы для опроса по теме 1

1. Дайте определение тензора. Приведите примеры тензоров
2. Операции над тензорами?
3. Как можно представить в виде тензоров временные ряды и изображения?
4. Понятие «глубокое обучение». Соотнесите между собой понятия «глубокое обучение», «машинное обучение», «искусственный интеллект».
5. Почему глубокое обучение получило развитие в последние годы?.
6. Модуль Numpy, Представление тензоров в данном модуле?

Назовите фрейморки, которые используются при построении нейронных сетей глубокого обучения.

7. Дайте общую характеристику языка Python, типы данных языка.

Тест

1. методы работы со списками

Поставьте соответствие между методами и результатами их применения, если исходный список имеет вид

ex=['a',2,'is',34.2]

1. ex.insert(3,2)
2. ex.append(2)
3. ex.append(2)
   ex.reverse()

Варианты ответа
[‘a’,2,’is’,2,34.2]
[‘a’,2,’is’, 34.2,2]
[2,34.2,’is’,2,’a’]

2. кортежи

Чем отличаются списки от кортежей?

1. ничем кроме названия и обозначений
2. кортеж изменяем, список -нет
3. кортеж не изменяем, список изменяем
4. элементы кортежа могут быть только одного типа
5. кортеж - это иерархический список

3. кортежи. Что из нижеперечисленного относится к кортежам?

1.(1,2,4,5)
2.[2,3,4]
3.{'a':1,'b':3}
4.'cortege'
5.1,2,3

4. Список. Какой список будет напечатан на экране при выполнении скрипта?

```
users = ["Tom", "Bob", "Alice", "Sam", "Bill"]
 users.sort()
users.reverse()
print(users)
```

1. ['Tom', 'Sam', 'Bob', 'Bill', 'Alice']
2. ['Alice','Bill','Bob','Sam','Tom']
3. ['Alice','Bill','Bob','Sam','Tom']
4. ['Bill','Sam','Alice','Bob','Tom']
5. будет отображаться пустой список

5.список. Что будет напечатано на экране в результате выполнения скрипта:

```
a = [1, 2, 4]
a[2] = 3
print(a)
```

1. 1,2,4
2. 1,2
3. 1,2,2
4. 1,2,2,3
5. 1,2,3

6.число элементов списка. Какие из перечисленных выражений создадут список ровно из трех элементов?

1. print('a b c'.split())
2. print(list(range(3)))
3. print('asd'.split())
4. print('a= ',1,2,3)
5. print('a= ',1,2,'3')

7. Индекс списка. Список задан перечислением элементов

```
#списки
numbers = [1, 2, 3, 4, 5,6,9]
```

print(numbers[-3])
Что будет напечатано на экране?

8.Срезы строки. Что вернет срез 'pandas'[-2:]. При ответе не забывайте указывать кавычки

9. срезы строки. Какой результат выражения 'numpy'[:3]*2? При ответе не забывать указывать одинарные кавычки

10.длина словаря. Чему равна длина словаря, полученного из списка

```
users_list = [
   ["+111123455", "Tom"],
   ["+384767557", "Bob"],
   ["+958758767", "Alice"]
]
users_dict = dict(users_list)
```

11. Метод объединения словарей. Имеются два словаря

```
d = {'a':1, 'b': 2}
d1={'c':'+79217071104','d':'+79263111423'}
```

Присоедините к первому словарю второй с помощью метода работы со словаря

Типовые оценочные материалы по теме 2

Задания по теме 2

Задание 1. Измеряются 13 характеристик химического состава вина. Необходимо по значениям имеющихся переменных определить тип вина.

Имеются данные о трех сортах вина. Сорт вина указан в трех столбцах класс_1, класс_2 и класс_3. Если в первом столбце (класс_3) стоит единица, то наблюдение соответствует виду вина третьего типа, если во втором столбце (класс_2) стоит единица, то наблюдение соответствует виду вина второго типа, если в третьем столбце (класс_1) стоит единица, то наблюдение соответствует виду вина первого типа. Таким образом, столбцы дублируют друг друга.

Список переменных:

1) Alcohol (содержание алкоголя)
2) Malic acid (яблочная кислота)
3) Ash (зола)
4) Alcalinity of ash
5) Magnesium (магний)
6) Total phenols (общее содержание фенола (карболовой кислоты))
7) Flavanoids (ароматические вещества)
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity (интенсивность цвета)
11) Hue (окраска? красители?)
12) OD280/OD315 of diluted wines
13) Proline (пролин)

Число наблюдений – 178

Число переменных – 13, все измерены в количественной (непрерывной) шкале.

Источник задачи: UCI Machine learning Database
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/wine
Обсуждение задачи:
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/wine

Задание 2. Имеются данные рейтинга Global Firepower, который основан на более чем 50 факторах, чтобы определить оценку PowerIndex (PwrIndx) данной страны. Данные взяты из наборов данных kaggle

Приведенная при расчете рейтинга формула позволяет более мелким, но более технологически развитым странам конкурировать с более крупными, менее развитыми. Модификаторы (в виде бонусов и штрафов) добавляются для дальнейшего уточнения списка. Некоторые пункты, которые соблюдены в отношении окончательного рейтинга:

- ранжирование не просто зависит от общего количества оружия, доступного какой-либо одной стране, а скорее сосредоточено на разнообразии оружия в пределах количества, чтобы обеспечить лучший баланс доступных огневых мощностей (т. е., например, 100 тральщиков не соответствует стратегической и тактической ценности 10 авианосцам);
- ядерные запасы не принимаются во внимание, но признанные/ подозреваемые ядерные державы получают бонус;
- географические факторы, логистическая гибкость, природные ресурсы и местная промышленность влияют на окончательный рейтинг;
- доступные трудовые ресурсы являются ключевым фактором; Страны с большим населением, как правило, выше;
- страны, не имеющие выхода к морю, «не наказаны» за отсутствие военно-морского флота; морские державы «наказываются» при расчете рейтинга из-за отсутствия разнообразия в имеющихся морских активах;
- альянсы НАТО получают небольшой бонус за теоретический обмен ресурсами, возможную организацию коалиций;
- нынешнее политическое/военное руководство, их политика, роль не принимается во внимание.

На 2017 год в базу данных GFP входит в общей сложности 133 страны. (http://www.globalfirepower.com/countries-listing.asp).

Решить задачу кластерного анализа методом иерархической кластеризации и методом k-средних. При решении задачи пропущенные данные заменять медианой. Для выявления пропущенных данных использовать библиотеку mice.

В анализируемом наборе имеется 47 атрибутов, первые два из которых символьные. Третий атрибут – ранг страны получается путем анализа всех остальных. Поэтому данный рейтинг не следует учитывать при решении задачи.

Задание 3. Даны данные учебного набора «Титаник» решить задачу классификации различными методами. Сравнить результаты классификации. Для решения задачи использовать файл train.csv. При решении задачи использовать следующие атрибуты

Pclass + Sex + Age + SibSp + Parch + Fare.

Для выделения нужных признаков использовать операцию конкатенации, например dat[,c(2,3,4,6,9)]

Для проверки качества классификации использовать тестовую и обучающую выборки. Размер выборок сделать равным. При построении выборки использовать функцию sample.

Построить таблицу сопряженности по результатам проверки качества работы классификатора на тестовой выборке.

Задание 4. Исследовать блокнот по предсказанию цены на недвижимость на основе набора данных boston

https://colab.research.google.com/drive/1p9CGM89OtVzN-gPFHpbfKmH-MItcuimD

Задание 5. Решить задачу классификации отзывов к фильмам с использованием dataset imdb. При решении задачи классификации набор данных взять в модуле keras. Задачу решить с помощью полносвязной нейронной сети, состоящей из двух уровней. Из 50000 отзывов на фильмы 25000 отзывов взять как обучающие и 25000 как контролирующие образцы.

Задание 5. Решить задачу классификации отзывов к фильмам с использованием dataset imdb. При решении задачи классификации набор данных взять в модуле keras. Задачу решить с помощью полносвязной нейронной сети, состоящей из двух уровней. Из 50000 отзывов на фильмы 25000 отзывов взять как обучающие и 25000 как контролирующие образцы.

Задание 6. Решить задачу классификации на основе набора данных Dataset of 11,228 newswires from Reuters, labeled over 46 topics.

Задание 7. Построить простую полносвязную нейронную сеть (feed forward neural network). Выходной слой с одним линейным нейроном — для задачи регрессии. Функция активации — RELU в промежуточном слое и sigmoid в выходном. При выполнении задания использовать пример: https://www.tensorflow.org/tutorials/keras/basic_regression.

Типовые вопросы для опроса по теме 2:

1. Дайте определение задачи классификации. Какие методы решения задачи классификации Вы знаете?
2. Особенности решения задач классификации с обучением.
3. Деревья решений и их свойства.
4. Ансамблевые методы машинного обучения.
5. Приведите примеры алгоритмов построения деревьев.
6. Как определяется правило остановки построения дерева?
7. Алгоритм CART? Приведите пример его использования.
8. Что понимается под кластером? Назовите характеристики кластера. Что такое «центроид» кластера?.
9. Дайте классификацию методов кластерного анализа. Приведите примеры их применения в практической жизни.
10. Зачем используются меры близости? Назовите методы определения близости между кластерами.
11. Когда применяется метод ближнего соседа, дальнего соседа? Сравните их.
12. Дайте характеристику метрик кластерного анализа.
13. Поясните содержание «дендограммы» и организацию ее применения.
14. Что понимается под профилем кластера.
15. Использование статистических пакетов для решения задач кластерного анализа.
16. Дайте характеристику метода к-средних
17. Сравните нейронные сети поверхностного и глубокого обучения.
18. Дайте общую характеристику сверточных нейронных сетей. Из каких слоев состоит такая сеть?
19. Операция свертывания. Шаблоны свертки. Какие основные операции используются в рекуррентных сетях?
20. Операция padding, maxpooling?
21. Организация обучения нейронной сети. Что такое batch? Сравните понятия batch, итерации, эпохи.
22. Что представляет собой проблема переобучения нейронной сети?

23. Что такое дообучение нейронной сети?

Тест

1. Выбор классификатора. Использование различных методов классификации позволяет получить следующие таблицы сопряженности.

Метод линейного дискриминантного анализа

69535	442
4546	477

Метод k ближайших соседей при k=9

69938	39
4973	50

Логистическая регрессия

69770	207
5005	18

Метод опорных векторов

69882	103
4890	125

Выбрать лучший метод по показателю precission

1. метод опорных векторов
2. метод k-ближайших соседей
3. логистическая регрессия
4. линейный дискриминантный анализ

2. Выбор классификатора. Использование различных методов классификации позволяет получить следующие таблицы сопряженности.

Метод линейного дискриминантного анализа

69535	442
.	477

Метод k ближайших соседей при k=9

69938	39
4973	50

Логистическая регрессия

69770	207
5005	18

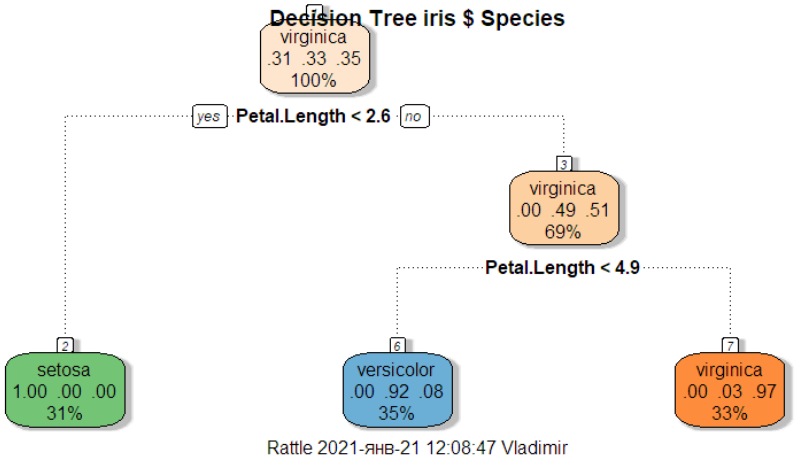
Метод опорных векторов

69882	103
4890	125

Выбрать лучший метод по показателю accuracy:

1) метод опорных векторов
2) метод k-ближайших соседей
3) логистическая регрессия
4) линейный дискриминантный анализ
5) деревья решений

3. При решении задачи классификации на наборе данных iris получено дерево решений



Какова вероятность ошибки при классификации цветка, если, длина лепестка больше 2, 6 см и меньше 4,9 см. Ответ дать с точностью до двух знаков после запятой. В качестве разделителя использовать запятую

4.Качество классификатора. Таблица сопряженности (матрица путаницы, confusion matrix), полученная при проверке качества бинарного классификатора, имеет вид

25	10
8	14

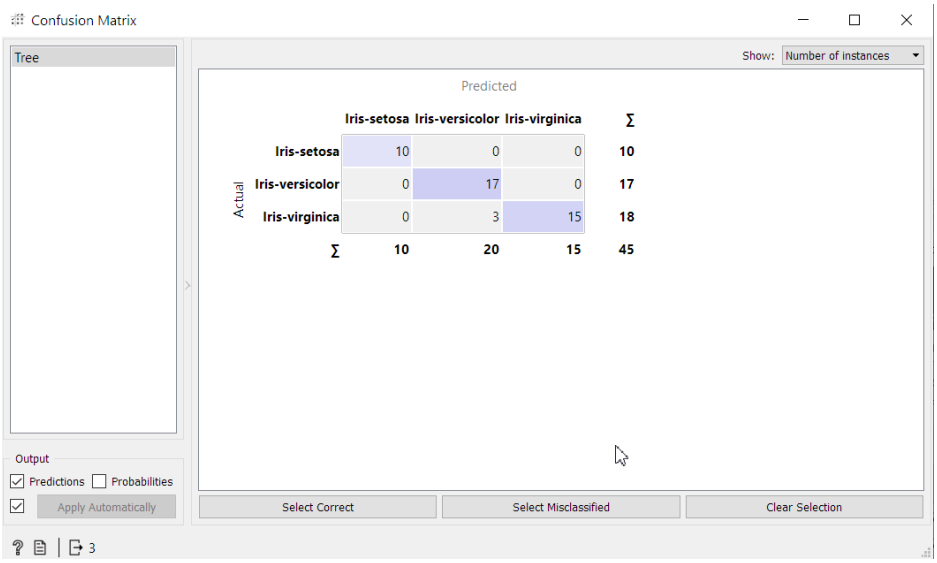
Строками матрицы являются истинные значения тестируемых объектов, а столбца - результаты тестирования. Чему равно значение показателя точности классификатора (precission)? Ответ дать с точностью до двух знаков после запятой

5. Качество классификатора. Таблица сопряженности (матрица путаницы, confusion matrix), полученная при проверке качества бинарного классификатора, имеет вид

25	10
8	14

Строками матрицы являются истинные значения тестируемых объектов, а столбца - результаты тестирования. Чему равна точность классификатора (recall)? Ответ дать с точностью до двух знаков после запятой

6. Качество классификации. В результате решения задачи классификации с помощью Orange методом деревьев решений получена матрица



Данная матрица получена при проверке качества классификатора с помощью тестовой выборки. Определить значение показателя AC. Ответ дать с точностью до двух знаков после запятой

7.Таблица сопряженности (confusion matrix)

Таблица сопряженности (матрица путаницы, confusion matrix), полученная при проверке качества бинарного классификатора, имеет вид

25	10
8	14

Строками матрицы являются истинные значения тестируемых объектов, а столбца - результаты тестирования. Чему равна точность классификатора (accuracy)? Ответ дать с точностью до двух знаков после запятой

Типовые оценочные материалы по теме 3
Задания по теме 3

Задание 1. Решить задачу классификации картинок на основе набора данных CIFAR-10 Датасет содержит 60000 цветных фото объектов 10 классов размером 32x32 пикселей.

https://colab.research.google.com/drive/1SLoQPhvv-MYllrNFPfv6y3y9eosubOrt

Задание 2. Решить задачу классификации картинок на основе набора данных **MNIST.** Нужно распознать рукописный текст по изображениям.

Задание 3. Используя свёрточную неройнную сеть YOLOv8 решить задачи обнаружения и детектирования элементов изображений на основе датасета

https://universe.roboflow.com/roboflow-100

Типовые вопросы для опроса по теме 3:

1. В чём заключается задача распознавания образов? Назовите основные методы представления данных для решения задач распознавания изображений.
2. Что такое детектирование объектов? Какие методы детектирования вы знаете?
3. Что такое сегментация изображений? Какие методы сегментации вы знаете?
4. Назовите основные методы обработки изображений.
5. Назовите основные методы фильтрации изображений.

6. В чём заключается синтез изображений?

7. Каким образом глубокое обучение используется в технологиях компьютерного зрения?

8. Назовите основные этапы построения сверхточной нейронной сети для решения задач анализа изображений.

Типовые оценочные материалы по теме 4
Задания по теме 4

Задание 1. Решить задачу анализа твиттов по тональности. Для решения задачи использовать наборы данных !wget https://www.dropbox.com/s/fnpq3z4bcnoktiv/positive.csv

!wget https://www.dropbox.com/s/r6u59ljhhjdg6j0/negative.csv

У нас есть выборка из твитов. Нам известна эмоциональная окраска каждого твита из выборки: положительная или отрицательная. Задача состоит в построении модели, которая по тексту твита предсказывает его эмоциональную окраску.

Классификацию по тональности используют в рекомендательных системах, чтобы понять, понравилось ли людям кафе, кино, etc.

При решении задачи использовать блокнот https://colab.research.google.com/drive/1zO1XlNcSFw6ZdtkXCGA909lP6OtZ4db1#scrollTo=KvOWdHcUWg98

Данный блокнот экпортировать в в файл блокнота *.ipynb. Открыть файл в блокноте notebook Anaconda navigator.

Задание 2. Самостоятельно рассмотреть блокнот на google colab. Экспортировать блокнот в notebook. Запустить notebook с помощью Anaconda navigator.

https://colab.research.google.com/drive/1SLoQPhvvMYllrNFPfv6y3y9eosubOrt#scrollTo=sjAzhUMjuwvi

Решить задачу анализа твиттов по тональности. Для решения задачи использовать наборы данных !wget https://www.dropbox.com/s/fnpq3z4bcnoktiv/positive.csv

!wget https://www.dropbox.com/s/r6u59ljhhjdg6j0/negative.csv

У нас есть выборка из твитов. Нам известна эмоциональная окраска каждого твита из выборки: положительная или отрицательная. Задача состоит в построении модели, которая по тексту твита предсказывает его эмоциональную окраску.

Классификацию по тональности используют в рекомендательных системах, чтобы понять, понравилось ли людям кафе, кино, etc.

Задание 4. С помощью блокнота google colab исследовать блокнот, в котором используется модель анализа текстов word2vec

Типовые вопросы для опроса по теме 4:

1. Особенности обработки текстов и последовательностей?
2. Как осуществляется обработка текста?
3. В чем отличие лемматизации и токенизации?
4. В чем отличие прямого кодирования и векторного представления слов?
5. Для чего используется слой embedding?
6. Дайте характеристику рекуррентных нейронных сетей
7. Дайте характеристику сетей LSTM

5. Оценочные средства для промежуточной аттестации по дисциплине

5.1. Зачёт проводится с использованием компьютерного теста с учетом результатов текущего контроля**.**

5.2. Оценочные средства для промежуточной аттестации

Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы. Показатели и критерии оценивания компетенций с учетом этапа их формирования

Таблица 5.1

Код компетенции	Наименование компетенции	Код Компонента компетенции	Наименование компонента компетенции
ПКС-2	Способен обосновывать подходы, используемые в бизнес-анализе, руководить и управлять бизнес-анализом с использованием информационно-коммуникационных технологий	ПКС-2.1	Способен использовать современные методы, информационные технологии, программный инструментарий в объеме, необходимом для решения задач бизнес-аналитики, использовать англоязычную документацию и справочные системы

Показатели и критерии оценивания компетенций на различных этапах их формирования

Таблица 5.2

Код компонента компетенции	**Промежуточный индикатор оценивания**	**Критерий оценивания**
ПКС-2.1	Использует современные методы, информационные технологии, программный инструментарий в объеме, необходимом для решения задач бизнес-аналитики, использовать англоязычную документацию и справочные системы	Использует информацию, методы и программные средства ее сбора, обработки и анализа, в том числе с использованием интеллектуальных методов. Самостоятельно решает задачи интеллектуального анализа данных, текстов и изображений с помощью аналитических платформ, фрейморков Python Keras, TensorFlow

Для оценки знаний и умений, соответствующих данным компетенциям, используются тестовые вопросы и задания.

Типовые вопросы, выносимые на зачет

1. Сравнительный анализ Python, R. Общая характеристика языка Python. Среда разработки. Платформа Anaconda.
2. Основы синтаксиса языка. Переменные, ключевые слова. Основы программирования на языке Python.
3. Типы данных. Функциональность для работы с данными. Установка пакетов

научных вычислений на Python. Установка Keras.

4. Основные методы работы с данными с использование Pandas.

5. Решение задач анализа данных. Понятие тензора. Скаляры, векторы, матрицы, тензоры третьего и высшего рангов. Временные ряды или последовательности. Изображения. Операции над тензорами.

6. Фреймворк Keras. Библиотека TensorFlow

7. Постановка задач кластерного анализа. Определение кластера. Параметры кластера. Меры близости. Метрики кластерного анализа.

8. Базовые алгоритмы кластеризации. Иерархическая кластеризация. Дендограммы.

9. Метод к-средних. Понятие центроида. Профили кластеров. Взаимосвязь кластерного и регрессионного анализа. Кластерный анализ на Python.

10. Формулировка задачи классификации. Классификационный анализ с обучением. Деревья решений. Алгоритмы построения деревьев решений. Методы и алгоритмы построения деревьев. Алгоритм CART. Определение прекращения построения дерева классификации.

11. Ансамблевые методы машинного обучения. Алгоритм Random Forest.

12. Кластеризация. Метод k-ближайших соседей.

13. Оценка качества задач классификации. Таблица сопряженности. Понятие чувствительности и специфичности. ROC-кривая. Ошибки первого и второго рода при решении задач классификации.

14. Понятие нейронной сети. Архитектура нейронной сети.

15. Основы машинного обучения. Оценка моделей машинного обучения. Тренировочные, проверочные и контрольные данные. Алгоритмы обучения. Функция потерь.

16. Механизм нейронных сетей на основе обучения. Обратное распространение ошибки.

17. Решение задач классификации, регрессии, прогнозирования с помощью нейронных сетей.

18. Понятие поверхностного и глубокого обучения. Глубокое обучение в технологиях компьютерного зрения.

19. Сверточные нейронные сети для решения задач распознавания образов.

20. Сверточные нейронные сети для решения задач детектирования и сегментации изображений.

21. Обработка и фильтрации изображений с использованием библиотек Python.

22. Глубокое обучение для текста и последовательностей. Рекуррентные нейронные сети. Слои LSTM, GRU.

23. Сверточные нейронные сети. Генеративное глубокое обучение.

Шкала оценивания

Оценка результатов производится на основе Положения о текущем контроле успеваемости обучающихся и промежуточной аттестации обегающихся по образовательным программам среднего профессионального и высшего образования в федеральном государственном бюджетном образовательном учреждении высшего образования «Российская академии народною хозяйства и государственной службы при Президенте Российской Федерации», утвержденного Приказом Ректора РАНХиГС при Президенте РФ от 30.01.2018 г. № 02-66 (п.10 раздела 3 (первый абзац) и п.11), а также Решения Ученого совета Северо-западного института управления РАНХиГС при Президенте РФ от 19.06.2018, протокол № 11.

Оценка «отлично» выставляется в случае, если при устном ответе студент проявил (показал):

- глубокое и системное знание всего программного материала учебного курса, изложил ответ последовательно и убедительно;

- отчетливое и свободное владение концептуально-понятийным аппаратом, научным языком и терминологией соответствующей дисциплины;
- умение правильно применять теоретические положения при решении практических вопросов и задач;
- умение самостоятельно выполнять предусмотренные программой задания;
- навык обоснования принятого решения.

Оценки «хорошо» выставляется в случае, если при устном ответе студент проявил (показал):
- знание узловых проблем программы и основного содержания лекционного курса;
- умение пользоваться концептуально-понятийным аппаратом умение преимущественно правильно применять теоретические положения при решении практических вопросов и задач,
- умение выполнять предусмотренные программой задания;
- в целом логически корректное, но не всегда точное и аргументированное изложение ответа.

Оценка «удовлетворительно» выставляется в случае, если при устном ответе студент проявил (показал):
- фрагментарные, поверхностные знания важнейших разделов программы и содержания лекционного курса;
- затруднения с использованием научно-понятийного аппарата и терминологии учебной дисциплины;
- затруднения с применением теоретических положений при решении практических вопросов и задач,

Оценка «неудовлетворительно» выставляется в случае, если при устном ответе студент проявил (показал):
- незнание либо отрывочное представление учебно-программного материала;
- неумение использовать научно-понятийный аппарат и терминологию учебной дисциплины;
- неумение применять теоретические положения при решении практических вопросов и задач,
- неумение выполнять предусмотренные программой задания.

6. Методические материалы по освоению дисциплины

Рабочей программой дисциплины предусмотрены следующие виды аудиторных занятий: лекции, практические занятия. На лекциях рассматриваются наиболее сложный материал дисциплины. Для развития у магистрантов креативного мышления и логики в каждой теме учебной дисциплины предусмотрены теоретические положения, инструментальные средства, а также примеры их использования при решении задач предиктивной аналитики. Кроме того, часть теоретического материала предоставляется на самостоятельное изучение по рекомендованным источникам для формирования навыка самообучения.

Практические занятия предназначены для самостоятельной работы магистрантов по решению конкретных задач. Каждое практическое занятие сопровождается заданиями, выдаваемыми магистрантам для решения во внеаудиторное время.

Для работы с печатными и электронными ресурсами СЗИУ имеется возможность доступа к электронным ресурсам. Организация работы магистрантов с электронной библиотекой указана на сайте института (странице сайта – «Научная библиотека»).

6.1. Методические указания для обучающихся по освоению дисциплины

Обучение по дисциплине «Интеллектуальны анализ данных» предполагает изучение курса на аудиторных занятиях (лекции, практические работы) и самостоятельной работы

обучающихся. Семинарские занятия дисциплины «Интеллектуальны анализ данных» предполагают их проведение в различных формах с целью выявления полученных знаний, умений, навыков и компетенций с проведением контрольных мероприятий. С целью обеспечения успешного обучения обучающийся должен готовиться к лекции, поскольку она является важнейшей формой организации учебного процесса, поскольку:

- знакомит с новым учебным материалом;
- разъясняет учебные элементы, трудные для понимания;
- систематизирует учебный материал;
- ориентирует в учебном процессе.

Подготовка к лекции заключается в следующем:

- внимательно прочитайте материал предыдущей лекции;
- узнайте тему предстоящей лекции (по тематическому плану, по информации лектора);
- ознакомьтесь с учебным материалом по рекомендуемой литературе;
- постарайтесь уяснить место изучаемой темы в своей профессиональной подготовке;
- запишите возможные вопросы, которые вы зададите лектору на лекции.

Подготовка к практическим занятиям:

- внимательно прочитайте материал лекций, относящихся к данному семинарскому занятию, ознакомьтесь с учебным материалом;
- ответьте на контрольные вопросы по семинарским занятиям, готовьтесь дать развернутый ответ на каждый из вопросов;
- уясните, какие учебные элементы остались для вас неясными и постарайтесь получить на них ответ заранее (до семинарского занятия) во время текущих консультаций преподавателя;
- готовиться можно индивидуально, парами или в составе малой группы, последние являются эффективными формами работы;
- рабочая программа дисциплины в части целей, перечню знаний, умений, терминов и учебных вопросов может быть использована вами в качестве ориентира в организации обучения.

Выполнение задания:

- выберите набор данных (временной ряд, временные ряды) для выполнения задания;
- выполните анализ используемых признаков (целевого признака);
- проанализируйте качество исходных данных;
- выполните выбор инструментов предобработки для улучшения качества исходных данных, а также формулировки предварительных гипотез;
- решите задачу прогнозирования уровней временного ряда;
- исследуйте возможность извлечения признаков временного ряда;
- решите задачу анализа выявленных признаков;
- оформите отчет по результатам выполнения задания.

7. Учебная литература и ресурсы информационно-телекоммуникационной сети "Интернет", включая перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

7.1. Основная 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.: ил. — (Серия «Бестселлеры O’Reilly»).

2. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. Пер. с англ. - СпБ.: ООО "Альфа-книга': 2018. - 688 с.: ил. - Парал. тит. англ.
3. Курносов М.Г. Введение в методы машинной обработки данных. - Новосибирск: Автограф. -220 с. Текст: электронный. - URL: http://www.iprbookshop.ru/102117.html (дата обращения: 11.01.2021). - Режим доступа: для авторизир. пользователей
4. Миркин, Борис Григорьевич. Введение в анализ данных – М.:Юрайт, 2020 – 174 с. Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/450262 (дата обращения: 01.10.2020)
5. Наумов В.Н. Анализ данных и машинное обучение: методы и инструментальные средства. Федер. гос. бюджет. образоват. учреждение высш. образования "Рос. акад. нар. хоз-ва и гос. службы при Президенте Рос. Федерации", Сев.-Зап. ин-т упр. - СПб. : СЗИУ - фил. РАНХиГС, 2020. - 260 с.
6. Сузи Р.А. Python. – СПб.: БХВ – Петербург, 2015 – 759 с. электронный ресурс
7. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.: ил.
8. Элбон К. Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. — СПб.: БХВ-Петербург, 2019. — 384 с.: ил.

Все источники основной литературы взаимозаменяемы.

7.2 Дополнительная литература

1. Мхитарян В. С., Архипова М. Ю., Дуброва Т. А., Миронкина Ю. Н., Сиротин В. П. Анализ данных. – М.: Юрайт, 2020 – 490 с. Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/450166 (дата обращения: 29.09.2020)
2. Нильсен Эйлин. Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение. –М.: ООО Диалектика – 2021 – 544 с.
3. Наумов, Владимир Николаевич. Средства бизнес- аналитики: учеб. пособие / В. Н. Наумов; Федер. гос. бюджет. образоват. учреждение высш. образования "Рос. акад. нар. хоз-ва и гос. службы при Президенте Рос. Федерации", Сев.-Зап. ин-т упр. - СПб. : СЗИУ - фил. РАНХиГС, 2016. - 107 с.
4. О`Нил К. Data Science: Инсайдерская информация для новичков. Включая язык R: [пер. с англ.] – СПб. Питер. – 368 с. Текст: электронный. - URL: http://new.ibooks.ru/bookshelf/359209/reading (дата обращения: 25.01.2021)
5. Шолле Ф. Глубокое обучение на Python. – СПб.: Питер. 2018. -400 с.
6. Шолле Ф. Глубокое обучение на R. – СПб.: Питер. 2018. -400 с.

7.3.Нормативные правовые документы

Приказ Минобрнауки России от 19.11.2013 N 1259 (ред. от 05.04.2016) "Об утверждении Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования - программам подготовки научно-педагогических кадров в аспирантуре (адъюнктуре)" (Зарегистрировано в Минюсте России 28.01.2014 N 31137)

7.4.Интернет-ресурсы

1. Электронно-образовательные ресурсы на сайте научной библиотеки СЗИУ РАНХиГС http://nwipa.ru
2. Электронные учебники электронно-библиотечной системы (ЭБС) «Айбукс» http://www.nwapa.spb.ru/index.php?page_id=76
   Электронные учебники электронно-библиотечной системы (ЭБС) «Лань» http://www.nwapa.spb.ru/index.php?page_id=76

3. Электронные учебники электронно-библиотечной системы (ЭБС) «IPRbooks» http://www.nwapa.spb.ru/index.php?page_id=76
4. Электронные учебники электронно-библиотечной системы (ЭБС) «Юрайт» http://www.nwapa.spb.ru/index.php?page_id=76
5. Научно-практические статьи по экономике и финансам Электронной библиотеки ИД «Гребенников» http://www.nwapa.spb.ru/index.php?page_id=76
6. Статьи из журналов и статистических изданий Ист-Вью http://www.nwapa.spb.ru/index.php?page_id=76
7. Англоязычные ресурсы EBSCO Publishing: доступ к мультидисциплинарным полнотекстовым базам данных различных мировых издательств по бизнесу, экономике, финансам, бухгалтерскому учету, гуманитарным и естественным областям знаний, рефератам и полным текстам публикаций из научных и научно–популярных журналов. Emerald eJournals Premier - крупнейшее мировое издательство, специализирующееся на электронных журналах и базах данных по экономике и менеджменту.
8. Социальная сеть специалистов по обработке данных и машинному обучению, система организации конкурсов по исследованию данных. http://www.kaggle.com
9. Социальная сеть специалистов по обработке данных и машинному обучению, система организации конкурсов по исследованию данных. http://www.roboflow.com
10. Интерактивная облачная среда программирования компании Google. https://colab.research.google.com
11. Официальный сайт среды программирования на языках Python и R, включающая набор популярных свободных библиотек, объединённых проблематиками науки о данных и машинного обучения Anaconda. https://www.anaconda.com

    Возможно использование, кроме вышеперечисленных ресурсов, и других электронных ресурсов сети Интернет.

7.5. Иные источники.

Не используются.

8. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы

Учебная дисциплина включает использование программного обеспечения Microsoft Excel, Microsoft Word, для подготовки текстового и табличного материала.

Интернет-сервисы и электронные ресурсы (поисковые системы, электронная почта, профессиональные тематические чаты и форумы, системы аудио и видео конференций, онлайн энциклопедии, справочники, библиотеки, электронные учебные и учебно-методические материалы).

2. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

№ п/п	Наименование
	Компьютерные классы с персональными ЭВМ, объединенными в локальные сети с выходом в Интернет
	Пакет Excel -2016 или новее, среда Anaconda
	Мультимедийные средства в каждом компьютерном классе и в лекционной аудитории

	Браузер, сетевые коммуникационные средства для выхода в Интернет

Компьютерные классы из расчета 1 ПЭВМ для одного обучаемого. Каждому обучающемуся должна быть предоставлена возможность доступа к сетям типа Интернет в течение не менее 20% времени, отведенного на самостоятельную подготовку.