

*Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»*

Факультет физико-математических и естественных наук

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины
Интеллектуальный анализ данных

Рекомендуется для направления подготовки

38.03.05 Бизнес-информатика

Квалификация (степень) выпускника
бакалавр

1. Цели и задачи дисциплины: Настоящий курс ставит своей целью ознакомление обучающихся с задачами, возникающими в области интеллектуального анализа данных (Data Mining), и методами их решения, которые помогут выявлять, формализовывать и успешно решать практические задачи интеллектуального анализа данных, возникающие в процессе профессиональной деятельности.

В ходе изучения дисциплины перед обучающимися ставятся следующие задачи:

- изучение методов и моделей интеллектуального анализа данных;
- получение представления об алгоритмах построения деревьев решений;
- изучение алгоритмов классификации и регрессии;
- изучение алгоритмов поиска ассоциативных правил;
- изучение методов кластеризации.

2. Место дисциплины в структуре ОП ВО:

Дисциплина «Интеллектуальный анализ данных» относится к *вариативной* части блока 1 учебного плана.

В таблице № 1 приведены предшествующие и последующие дисциплины, направленные на формирование компетенций дисциплины в соответствии с матрицей компетенций ОС ВО.

Таблица № 1

Предшествующие и последующие дисциплины, направленные на формирование компетенций

№ п/п	Шифр и наименование компетенции	Предшествующие дисциплины	Последующие дисциплины (группы дисциплин)
Универсальные компетенции			
1.	УК-1, УК-12	Анализ данных, Архитектура предприятия; Основы разработки корпоративных инфокоммуникационных систем; Системы поддержки принятия решений, Методы машинного обучения	-
Общепрофессиональные компетенции			
2.	ОПК-2, ОПК-6, ОПК-7	Основы программирования, Технология программирования, Алгоритмы и анализ сложности, Java и его приложения, Архитектура предприятия; Основы разработки корпоративных инфокоммуникационных систем; Анализ данных, Системы поддержки принятия решений, Методы машинного обучения	-
Профессиональные компетенции (вид профессиональной деятельности: <i>научно-исследовательская</i>)			
1.	ПК-3, ПК-4	Основы программирования Технология программирования, Алгоритмы и анализ сложности, Java и его приложения, Анализ данных, Методы машинного обучения	-

Профессионально-специализированные компетенции специализации			
-	-	-	-

3. Требования к результатам освоения дисциплины:

Процесс изучения дисциплины направлен на формирование следующих компетенций:
УК-1; УК-12; ОПК-2; ОПК-6; ОПК-7; ПК-3; ПК-4

УК-1: Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач:

- УК-1.1 Знает принципы сбора, отбора и обобщения информации, методики системного подхода для решения профессиональных задач
- УК-1.2 Умеет анализировать и систематизировать разнородные данные, оценивать эффективность процедур анализа проблем и принятия решений в профессиональной деятельности
- УК-1.3 Владеет навыками научного поиска и практической работы с информационными источниками; методами принятия решений

УК-12 Способен: искать нужные источники информации и данные, воспринимать, анализировать, запоминать и передавать информацию с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач; проводить оценку информации, ее достоверность, строить логические умозаключения на основании поступающих информации и данных

УК-12.1 Способен: искать нужные источники информации и данные, воспринимать, анализировать, запоминать и передавать информацию с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач; проводить оценку информации, ее достоверность, строить логические умозаключения на основании поступающих информации и данных

ОПК-2 Способен проводить исследование и анализ рынка информационных систем и информационно-коммуникационных технологий, выбирать рациональные решения для управления бизнесом

- ОПК-2.1 Знает методы сбора, анализа, систематизации, хранения и поддержания в актуальном состоянии информации для проведения исследования рынка ИС и ИКТ
- ОПК-2.2 Умеет проводить сбор, анализ, систематизацию информации для проведения исследования рынка ИС и ИКТ

ОПК-6 Способен выполнять отдельные задачи в рамках коллективной научно-исследовательской, проектной и учебно-профессиональной деятельности для поиска, выработки и применения новых решений в области информационно-коммуникационных технологий

- ОПК-6.1 Способен выполнять отдельные задачи в рамках коллективной научно-исследовательской, проектной и учебно-профессиональной деятельности для поиска, выработки и применения новых решений в области информационно-коммуникационных технологий

ОПК-7 Способен использовать цифровые технологии и методы в профессиональной деятельности в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

- ОПК-7.1 Знает базовые принципы цифровых технологий и методов, необходимых в профессиональной деятельности в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

- ОПК-7.2 Умеет применять необходимые в профессиональной деятельности цифровые технологии и методы в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.
- ОПК-7.3 Владеет необходимыми в профессиональной деятельности технологиями и методами в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

ПК-3 Способен выполнять работы и управлять работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы

- ПК-3.3 Знает основы программирования; современные объектно-ориентированные языки программирования; современные структурные языки программирования; языки современных бизнес-приложений
- ПК-3.4 Умеет кодировать на языках программирования; тестировать результаты кодирования

ПК-4 Способен принимать обоснованные управленческие решения в своей профессиональной деятельности

- ПК-4.2 Умеет анализировать и оценивать факторы и условия, влияющие на принятие управленческих решений
- ПК-4.3 Умеет проводить оценку эффективности принятия решения в соответствии с выбранными критериями или выбранными целевыми показателями

В результате изучения дисциплины студент должен:

Знать: принципы обработки больших массивов данных, способы их представления и хранения; основные задачи и методы интеллектуального анализа данных; возможности современных и перспективных средств разработки программных продуктов, технических средств.

Уметь: формулировать задачи анализа данных; выбирать адекватные алгоритмы их решения; выполнять процедуры проектирования хранилищ данных и заполнения готовых хранилищ данными; оценивать качество получаемых решений; выбирать средства реализации требований к программному обеспечению.

Владеть: технологиями разработки алгоритмов и программными системами анализа данных; средствами автоматизации интеллектуального анализа и обработки данных; формированием и предоставлением отчетности в соответствии с установленными регламентами.

4. Объем дисциплины и виды учебной работы

Общая трудоемкость дисциплины составляет 4 зачетные единицы.

Вид учебной работы	Всего часов	Семестры
		Семестр 7, модуль 13 (D)
Аудиторные занятия (всего)	36	36
<i>Лекции</i>	18	18
<i>Практические занятия (ПЗ)</i>	18	18
<i>Семинары (С)</i>	-	-
<i>Лабораторные работы (ЛР)</i>	-	-
Самостоятельная работа (всего)	108	108
Общая трудоемкость, час.	144	144
зач. ед.	4	4

5. Содержание дисциплины

5.1. Содержание разделов дисциплины

№ п/п	Наименование раздела дисциплины	Содержание раздела (темы)
1.	Введение в интеллектуальный анализ данных	Тема 1. Интеллектуальный анализ данных. Матрица данных. Набор данных «Ирисы». Числовые и категориальные признаки. Случайная величина. Случайная выборка. Эмпирическая функция распределения. Ковариация. Корреляция. Ковариационная матрица. Основные этапы интеллектуального анализа данных. Библиотеки языка Python для интеллектуального анализа данных. Репозиторий данных машинного обучения UCI.
2.	Подготовка и обработка данных. Снижение размерности	Тема 1. Предварительная обработка данных. Очистка данных. Пропущенные значения. Зашумленные данные. Нормализация данных. Снижение размерности данных. Проекция в подпространство. Вектор ошибки. Метод главных компонент. Направление с максимальной дисперсией. Минимальная квадратичная ошибка. Алгоритм PCA. Главные компоненты набора «Ирисы».
3.	Поиск ассоциативных правил	Тема 1. Наборы объектов. Представление базы данных. Бинарная база данных. Поддержка и частые наборы объектов. Ассоциативные правила. Интеллектуальный анализ наборов объектов и правил. Алгоритм Apriori. Алгоритмы Eclat, dEclat, FPGrowth. Алгоритм для ассоциативных правил.
4.	Кластеризация данных	Тема 1. Постановка задачи кластеризации данных. Алгоритм К-средних. Автоматический выбор начальных центров кластеров. Иерархическая кластеризация. Метрики качества кластеризации.
5.	Классификация данных	Тема 1. Постановка задачи классификации. Метрики качества классификации. Байесовская классификация. Классификация по ближайшим соседям. Вероятность ошибки при классификации по ближайшему соседу. Тема 2. Регрессионный анализ. Линейный дискриминантный анализ. Алгоритм линейного дискриминантного анализа. Ядерный дискриминантный анализ. Алгоритм ядерного дискриминантного анализа. Тема 3. Обучение дерева решений. Классификатор дерева решений. Точки разбиения. Разбиение данных и чистота. Алгоритм построения дерева решений. Энтропия. Индекс Джини. Мера CART. Алгоритм оценки числовых атрибутов. Алгоритм оценки категориальных атрибутов. Тема 4. Метод опорных векторов (SVM). Гиперплоскости. Разделяющая гиперплоскость. Маржа и опорные вектора. Целевая функция и линейные ограничения. Классификатор SVM. Нелинейный случай. Алгоритмы обучения SVM. Градиентный спуск. Стохастический градиентный спуск.

5.2. Разделы дисциплин и виды занятий

№ п/п	Наименование раздела дисциплины	Лек-ции	Прак-зан.	Лаб. зан.	Сем. зан.	СР С	Всего час.
1.	Введение в интеллектуальный анализ данных	2	2	-	-	12	16
2.	Подготовка и обработка данных. Снижение размерности.	2	2	-	-	12	16
3.	Поиск ассоциативных правил	2	2	-	-	12	16
4.	Кластеризация данных	2	2	-	-	12	16
5.	Классификация данных	10	10	-	-	60	80
	ВСЕГО:	18	18	-	-	108	144

6. Лабораторные работы не предусмотрены

7. Практические занятия

№ п/п	№ раздела	Наименование практических занятий	Трудо-емкость (час.)
1.	1	1. Расчет статистических показателей и визуализация заданного набора данных	2
2.	2	1. Использование метода главных компонент для снижения размерности данных.	2
3.	3	1. Поиск ассоциативных правил при помощи алгоритмов Apriori, Eclat, Declat, FPGrowth	2
4.	4	1. Решение задачи кластеризации данных при помощи алгоритма К-средних и иерархической кластеризации.	2
5.	5	1. Решение задачи классификации данных при помощи байесовской классификации и классификации по ближайшим соседям. 2. Решение задач классификации и прогнозирования данных при помощи регрессионного анализа. 3. Решение задачи классификации данных при помощи обучения дерева решений. 4. Решение задачи классификации данных при помощи метода опорных векторов.	10

8. Материально-техническое обеспечение дисциплины:

Учебная аудитория с ПК и проектором для проведения учебных занятий (в том числе для практического и лекционного типов занятий, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации).

Компьютерные (дисплейные) классы с доступом к сети Интернет и электронно-образовательной среде Университета для выполнения обучающимися заданий лабораторного практикума, выполнения обучающимися самостоятельной работы и проведения компьютерного тестирования обучающихся (при необходимости).

9. Информационное обеспечение дисциплины

а) программное обеспечение

- ОС Windows, MS Office (программа корпоративного лицензирования (Microsoft Subscription) Enrollment for Education Solutions), браузер Firefox (лицензия MPL-

2.0) или браузер Chrome (лицензия Google Chrome Terms of Service); Adobe Reader (Adobe Software License Agreement)

- программное обеспечение со свободной лицензией: Dev-C++ (лицензия GNU GPL)
- ОС Linux, офисный пакет LibreOffice (лицензия MPL-2.0), ПО для просмотра pdf (например, evince (лицензия GPL-2+ CC-BY-SA-3.0)), dev-lang/perl (лицензия Artistic GPL-1+), dev-lang/python (лицензия PSF-2), numpy (лицензия NumPy license), sympy (лицензия The 3-Clause BSD License), matplotlib (лицензия Python Software Foundation License), dev-lang/R (лицензия || (GPL-2 GPL-3) LGPL-2.1)

б) базы данных, информационно-справочные и поисковые системы:

2. Сайт библиотеки РУДН <http://lib.rudn.ru/>
3. ТУИС <http://esystem.pfur.ru/>

10. Учебно-методическое обеспечение дисциплины:

а) основная литература:

1. Data mining // [Электронный ресурс] URL: <https://www.intuit.ru/studies/courses/6/6/info>, режим доступа: свободный.

б) дополнительная литература:

1. Введение в аналитику больших массивов данных // [Электронный ресурс] URL: <https://www.intuit.ru/studies/courses/12385/1181/info>, режим доступа: свободный.

11. Методические указания для обучающихся по освоению дисциплины

Учебным планом на изучение дисциплины отводится 1 семестр. В течение семестра выполняются лабораторные работы, домашние задания и проводятся контрольные мероприятия. В конце семестра производится итоговый контроль знаний - экзамен.

11.1 Методические указания по самостоятельному освоению теоретического материала по дисциплине

Лекционный материал дисциплины охватывает темы, указанные в разделе 5.1 программы дисциплины. В ТУИС (<http://esystem.pfur.ru>) по темам лекций размещены учебные материалы. Рекомендуется по указанным темам в дополнение к учебным материалам ТУИС изучить материалы, указанные в п. 10 программы дисциплины.

11.2 Методические указания по выполнению домашних работ

Задания по домашним работам выполняются индивидуально каждым студентом в соответствии с календарным планом. По результатам выполнения каждого домашнего задания студентом готовится отчет.

11.3. Методические указания по подготовке к контрольным мероприятиям

Контрольные мероприятия по дисциплине проводятся в форме контрольных работ и/или оценки результатов выполнения домашних заданий. Итоговый контроль в форме опроса проводится по темам всех разделов дисциплины. Материалы для подготовки к промежуточному и итоговому контролю размещены в соответствующем разделе ТУИС (<http://esystem.pfur.ru>).

12. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

ФОС по дисциплине представлен в приложении к данной программе.

Программа составлена в соответствии с требованиями ОС ВО РУДН.

Разработчик:

доцент кафедры
информационных технологий, к.ф.-м.н.



С.Г. Шорохов

Зав. кафедрой информационных
технологий, д.ф.-м.н.



Ю.Н. Орлов

Руководитель программы

Заведующий кафедрой
прикладной информатики
и теории вероятностей, д.т.н., проф.



К.Е. Самуйлов

*Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»*

Факультет физико-математических и естественных наук

Кафедра информационных технологий

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

ПО УЧЕБНОЙ ДИСЦИПЛИНЕ
Интеллектуальный анализ данных

Рекомендуется для направления подготовки
38.03.05 «Бизнес-информатика»

Квалификация (степень) выпускника
Бакалавр

Паспорт фонда оценочных средств по дисциплине

Дисциплина: Интеллектуальный анализ данных

Направление: 38.03.05 «Бизнес-информатика»

Код контролируемой компетенции или ее	Контролируемый раздел дисциплины	Контролируемая тема дисциплины	ФОСы (формы контроля уровня освоения ООП)			Баллы темы	Баллы раздела
			Аудиторная работа	Самостоятельная работа	Экзамен/Зачет		
			Опрос	Выполнение ДЗ			
УК-1; УК-12; ОПК-2; ОПК-6; ОПК-7; ПК-3; ПК-4	Раздел 1: Введение в интеллектуальный анализ данных	Тема 1. Интеллектуальный анализ данных. Матрица данных. Набор данных «Ирисы». Числовые и категориальные признаки. Случайная величина. Случайная выборка. Эмпирическая функция распределения. Ковариация. Корреляция. Ковариационная матрица. Основные этапы интеллектуального анализа данных. Библиотеки языка Python для интеллектуального анализа данных. Репозиторий данных машинного обучения UCI.		10	2	12	12
	Раздел 2: Подготовка и обработка данных. Снижение размерности	Тема 1. Предварительная обработка данных. Очистка данных. Пропущенные значения. Зашумленные данные. Нормализация данных. Снижение размерности данных. Проекция в подпространство. Вектор ошибки. Метод главных компонент. Направление с максимальной дисперсией. Минимальная квадратичная ошибка. Алгоритм PCA. Главные компоненты набора «Ирисы».		10	2	12	12
	Раздел 3: Поиск ассоциативных правил	Тема 1. Наборы объектов. Представление базы данных. Бинарная база данных. Поддержка и частые наборы объектов. Ассоциативные правила. Интеллектуальный анализ наборов объектов и правил. Алгоритм Apriori. Алгоритмы Eclat, dEclat, FPGrowth. Алгоритм для ассоциативных правил.		10	2	12	12
	Раздел 4: Кластеризация данных	Тема 1. Постановка задачи кластеризации данных. Алгоритм K-средних. Автоматический выбор начальных центров кластеров. Иерархическая кластеризация. Метрики качества кластеризации.		10	2	12	12

	Раздел 5: Классификация данных	Тема 1. Постановка задачи классификации. Метрики качества классификации. Байесовская классификация. Классификация по ближайшим соседям. Вероятность ошибки при классификации по ближайшему соседу.		10	3	13	52
		Тема 2. Регрессионный анализ. Линейный дискриминантный анализ. Алгоритм линейного дискриминантного анализа. Ядерный дискриминантный анализ. Алгоритм ядерного дискриминантного анализа.		10	3	13	
		Тема 3. Обучение дерева решений. Классификатор дерева решений. Точки разбиения. Разбиение данных и чистота. Алгоритм построения дерева решений. Энтропия. Индекс Джини. Мера CART. Алгоритм оценки числовых атрибутов. Алгоритм оценки категориальных атрибутов.		10	3	13	
		Тема 4. Метод опорных векторов (SVM). Гиперплоскости. Разделяющая гиперплоскость. Маржа и опорные вектора. Целевая функция и линейные ограничения. Классификатор SVM. Нелинейный случай. Алгоритмы обучения SVM. Градиентный спуск. Стохастический градиентный спуск.		10	3	13	
		ИТОГО:		80	20	100	100

Процесс изучения дисциплины направлен на формирование следующих компетенций:

УК-1; УК-12; ОПК-2; ОПК-6; ОПК-7; ПК-3; ПК-4

УК-1: Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач:

- УК-1.1 Знает принципы сбора, отбора и обобщения информации, методики системного подхода для решения профессиональных задач
- УК-1.2 Умеет анализировать и систематизировать разнородные данные, оценивать эффективность процедур анализа проблем и принятия решений в профессиональной деятельности
- УК-1.3 Владеет навыками научного поиска и практической работы с информационными источниками; методами принятия решений

УК-12 Способен: искать нужные источники информации и данные, воспринимать, анализировать, запоминать и передавать информацию с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач; проводить оценку информации, ее достоверность, строить логические умозаключения на основании поступающих информации и данных

УК-12.1 Способен: искать нужные источники информации и данные, воспринимать, анализировать, запоминать и передавать информацию с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач; проводить оценку информации, ее достоверность, строить логические умозаключения на основании поступающих информации и данных

ОПК-2 Способен проводить исследование и анализ рынка информационных систем и информационно-коммуникационных технологий, выбирать рациональные решения для управления бизнесом

- ОПК-2.1 Знает методы сбора, анализа, систематизации, хранения и поддержания в актуальном состоянии информации для проведения исследования рынка ИС и ИКТ

- ОПК-2.2 Умеет проводить сбор, анализ, систематизацию информации для проведения исследования рынка ИС и ИКТ

ОПК-6 Способен выполнять отдельные задачи в рамках коллективной научно-исследовательской, проектной и учебно-профессиональной деятельности для поиска, выработки и применения новых решений в области информационно коммуникационных технологий

- ОПК-6.1 Способен выполнять отдельные задачи в рамках коллективной научно-исследовательской, проектной и учебно-профессиональной деятельности для поиска, выработки и применения новых решений в области информационно-коммуникационных технологий

ОПК-7 Способен использовать цифровые технологии и методы в профессиональной деятельности в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

- ОПК-7.1 Знает базовые принципы цифровых технологий и методов, необходимых в профессиональной деятельности в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

- ОПК-7.2 Умеет применять необходимые в профессиональной деятельности цифровые технологии и методы в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

- ОПК-7.3 Владеет необходимыми в профессиональной деятельности технологиями и методами в области бизнес-информатики для: изучения и моделирования объектов профессиональной деятельности, анализа данных, представления информации и пр.

ПК-3 Способен выполнять работы и управлять работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы

- ПК-3.3 Знает основы программирования; современные объектно-ориентированные языки программирования; современные структурные языки программирования; языки современных бизнес-приложений

- ПК-3.4 Умеет кодировать на языках программирования; тестировать результаты кодирования

ПК-4 Способен принимать обоснованные управленческие решения в своей профессиональной деятельности

- ПК-4.2 Умеет анализировать и оценивать факторы и условия, влияющие на принятие управленческих решений

- ПК-4.3 Умеет проводить оценку эффективности принятия решения в соответствии с выбранными критериями или выбранными целевыми показателями

Балльно-рейтинговая система оценки уровня знаний

Таблица соответствия баллов и оценок

Баллы БРС	Традиционные оценки РФ	Оценки ECTS
95 - 100	5	A
86 - 94		B
69 - 85	4	C
61 - 68	3	D
51 - 60		E
31 - 50	2	FX
0 - 30		F
51-100	Зачет	Passed

Правила применения БРС

1. Раздел (тема) учебной дисциплины считаются освоенными, если студент набрал более 50 % от возможного числа баллов по этому разделу (теме).
2. Студент не может быть аттестован по дисциплине, если он не освоил все темы и разделы дисциплины, указанные в сводной оценочной таблице дисциплины.
3. По решению преподавателя и с согласия студентов, не освоивших отдельные разделы (темы) изучаемой дисциплины, в течение учебного семестра могут быть повторно проведены мероприятия текущего контроля успеваемости или выданы дополнительные учебные задания по этим темам или разделам. При этом студентам за данную работу засчитывается минимально возможный положительный балл (51 % от максимального балла).
4. При выполнении студентом дополнительных учебных заданий или повторного прохождения мероприятий текущего контроля полученные им баллы засчитываются за конкретные темы. Итоговая сумма баллов не может превышать максимального количества баллов, установленного по данным темам (в соответствии с приказом Ректора № 564 от 20.06.2013). По решению преподавателя предыдущие баллы, полученные студентом по учебным заданиям, могут быть аннулированы.
5. График проведения мероприятий текущего контроля успеваемости формируется в соответствии с календарным планом курса. Студенты обязаны сдавать все задания в сроки, установленные преподавателем.
6. Время, которое отводится студенту на выполнение мероприятий текущего контроля успеваемости, устанавливается преподавателем. По завершение отведенного времени студент должен сдать работу преподавателю, вне зависимости от того, завершена она или нет.
7. Использование источников (в том числе конспектов лекций и практических заданий) во время выполнения контрольных мероприятий возможно только с разрешения преподавателя.
8. Отсрочка в прохождении мероприятий текущего контроля успеваемости считается уважительной только в случае болезни студента, что подтверждается наличием у него медицинской справки, заверенной круглой

печатью в поликлинике № 25, предоставляемой преподавателю не позднее двух недель после выздоровления. В этом случае выполнение контрольных мероприятий осуществляется после выздоровления студента в срок, назначенный преподавателем. В противном случае, отсутствие студента на контрольном мероприятии признается не уважительным.

9. Студент допускается к итоговому контролю знаний с любым количеством баллов, набранных в семестре.
10. Итоговый контроль знаний оценивается из 20 баллов независимо от числа баллов за семестр.
11. Если в итоге за семестр студент получил менее 31 балла, то ему выставляется оценка F и студент должен повторить эту дисциплину в установленном порядке. Если же в итоге студент получил 31-50 баллов (т. е. FX), то студенту разрешается добор необходимого (до 51) количества баллов путем повторного одноразового выполнения предусмотренных контрольных мероприятий, при этом по усмотрению преподавателя аннулируются соответствующие предыдущие результаты. Ликвидация задолженностей проводится в период с 07.02 по 28.02 (с 07.09 по 28.09) по согласованию с деканатом.

Примерный перечень оценочных средств

п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в фонде
<i>Аудиторная работа</i>			
1			
3	Опрос *	Средство контроля, организованное как специальная беседа преподавателя с обучающимся на темы, связанные с изучаемой дисциплиной, и рассчитанное на выяснение объема знаний обучающегося по определенному разделу или теме.	Вопросы по темам/разделам дисциплины
4	Дифф. зач. *	Оценка работы студента в течение семестра (года, всего срока обучения и др.) и призван выявить уровень, прочность и систематичность полученных им теоретических и практических знаний, приобретения навыков самостоятельной работы, развития творческого мышления, умение синтезировать полученные знания и применять их в решении практических задач.	Примеры заданий/вопросов, пример билета
<i>Самостоятельная работа</i>			
1	Домашняя работа	Система практических заданий, направленных на формирование практических навыков у обучающихся	Фонд практических заданий

Учебным планом на изучение дисциплины отводится один семестр. В дисциплине предусмотрены лекции, практические занятия, контрольные мероприятия по проверке отчётов по домашним работам. В конце семестра проводится итоговый контроль знаний.

Оценивание результатов освоения дисциплины производится в соответствии с балльно-рейтинговой системой. По дисциплине предусмотрен экзамен.

(*) Итоговый контроль знаний по дисциплине проводится в форме письменного ответа на вопросы из билетов или в форме опроса.

Критерии оценки по дисциплине

95-100 баллов:

- полное и своевременное выполнение на высоком уровне практических заданий с оформлением отчетов, успешное прохождение контрольных мероприятий, предусмотренных программой курса;
- систематизированное, глубокое и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- использование научной терминологии, стилистически грамотное, логически правильное изложение ответов на вопросы, умение делать обоснованные выводы;
- безупречное владение программным обеспечением, умение эффективно использовать его в постановке и решении научных и профессиональных задач;
- выраженная способность самостоятельно и творчески решать поставленные задачи;
- полная самостоятельность и творческий подход при изложении материала по программе дисциплины;
- полное и глубокое усвоение основной и дополнительной литературы, рекомендованной программой дисциплины и преподавателем.

86- 94 балла:

- полное и своевременное выполнение на хорошем уровне практических заданий с оформлением отчетов, успешное прохождение контрольных мероприятий, предусмотренных программой курса;
- систематизированное, глубокое и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- использование научной терминологии, стилистически грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы;
- хорошее владение программным обеспечением, умение эффективно использовать его в постановке и решении научных и профессиональных задач;
- способность самостоятельно решать поставленные задачи в нестандартных производственных ситуациях;
- усвоение основной и дополнительной литературы, нормативных и законодательных актов, рекомендованных программой дисциплины и преподавателем.

69-85 баллов:

- своевременное выполнение на хорошем уровне практических заданий с оформлением отчетов, прохождение контрольных мероприятий, предусмотренных программой курса;
- хороший уровень культуры исполнения практических заданий;
- систематизированное и полное освоение навыков и компетенций по всем разделам программы дисциплины;

- владение программным обеспечением, умение использовать его в постановке и решении научных и профессиональных задач;
- способность самостоятельно решать проблемы в рамках программы дисциплины;
- усвоение основной литературы;

51-68 баллов:

- выполнение на удовлетворительном уровне практических заданий с оформлением отчетов, прохождение контрольных мероприятий, предусмотренных программой курса;
- систематизированное и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- удовлетворительное владение программным обеспечением, умение использовать его в постановке и решении научных и профессиональных задач;
- способность решать проблемы в рамках программы дисциплины;
- удовлетворительное усвоение основной литературы;

31 - 50 баллов – НЕ ЗАЧТЕНО:

- не выполнение, несвоевременное выполнение или выполнение на неудовлетворительном уровне практических заданий, не прохождение контрольных мероприятий, предусмотренных программой курса;
- недостаточно полный объем навыков и компетенции в рамках программы дисциплины;
- неумение использовать в практической деятельности научной терминологии, изложение ответа на вопросы с существенными стилистическими и логическими ошибками;
- слабое владение программным обеспечением по разделам программы дисциплины, некомпетентность в решении стандартных (типовых) производственных задач;
- способность решать проблемы в рамках программы дисциплины;
- удовлетворительное усвоение основной литературы;

0-30 баллов, НЕ ЗАЧТЕНО:

- отсутствие умений, навыков, знаний и компетенции в рамках программы дисциплины;
- невыполнение домашних заданий, не прохождение контрольных мероприятий, предусмотренных программой курса; отказ от ответов по программе дисциплины;
- игнорирование занятий по дисциплине по неуважительной причине.

Вопросы для итогового контроля знаний

Дисциплина Интеллектуальный анализ данных

1. Предварительная обработка данных. Пропущенные значения. Зашумленные данные. Метод биннинга. Преобразование данных. Нормализация данных.
2. Снижение размерности данных. Проекция в подпространство. Вектор ошибки.
3. Метод главных компонент (PCA). Направление с максимальной дисперсией. Минимальная среднеквадратичная ошибка (MSE).
4. Метод главных компонент (PCA). Наилучшая g -мерная аппроксимация. Выбор размерности.
5. Алгоритм метода главных компонент (PCA). Геометрия метода главных компонент.
6. Задача поиска ассоциативных правил. Наборы предметов. Транзакции. Бинарная база данных. Горизонтальное и вертикальное представления.
7. Поддержка набора предметов. Популярные наборы предметов. Ассоциативное правило. Поддержка правила. Достоверность правила. Популярные и сильные правила.
8. Майнинг наборов предметов и правил. Алгоритм Brute Force. Пространство и дерево поиска наборов предметов.
9. Поиск наборов предметов: алгоритм Apriori. Пример.
10. Поиск наборов предметов: алгоритм Eclat. Пересечения наборов транзакций. Пример.
11. Поиск наборов предметов: алгоритм dEclat. Разности наборов транзакций. Пример.
12. Алгоритм построения ассоциативных правил. Пример.
13. Постановка задачи кластеризации. Метрики для ошибки кластеризации. Кластеризация методом полного перебора.
14. Кластеризация через представителей. Алгоритм k средних (k -means). Сходимость алгоритма. Пример.
15. Начальные центры кластеров. Алгоритм выбора начальных центров кластеров.
16. Иерархическая кластеризация. Вложенные разбиения. Дендрограмма кластеризации.
17. Агломеративная иерархическая кластеризация. Алгоритм агломеративной кластеризации.
18. Расстояние между кластерами. Кластеризация методом одиночной связи. Формула Ланса-Уильямса.
19. Меры качества кластеризации. Таблица сопряженности. Чистота кластера. Чистота кластеризации.
20. Байесовский классификатор. Теорема Байеса. Оценка априорной вероятности класса.
21. Байесовский классификатор. Параметрический подход для числовых признаков. Алгоритм байесовской классификации. Пример.
22. Байесовский классификатор. Классификация категориальных признаков. Пример.
23. Наивный байесовский классификатор. Алгоритм наивной байесовской классификации. Пример.
24. Наивный байесовский классификатор. Классификация категориальных признаков. Пример.
25. Метод K ближайших соседей. Пример.
26. Классификатор дерева решений. Рекурсивные разбиения. Гиперплоскости. Чистота области.
27. Алгоритм построения дерева принятия решений.
28. Оценка разбиения: энтропия, информационный выигрыш, индекс Джини.
29. Оценка разбиения для числовых признаков. Алгоритм оценки числовых признаков.

30. Оценка разбиения для категориальных признаков. Алгоритм оценки категориальных признаков.
31. Линейный дискриминантный анализ. Проекция на прямую. Оптимальный линейный дискриминант.
32. Линейный дискриминантный анализ. Линейный дискриминант Фишера. Алгоритм линейного дискриминанта.
33. Оценка классификации. Меры качества классификации: доля ошибок, точность.
34. Меры оценки качества классификации на основе таблицы сопряженности. Точность и полнота класса. F-мера.
35. Регрессионная модель. Линейная регрессия. Одномерный случай. Нелинейная регрессия.
36. Оценка качества регрессии. Коэффициент детерминации.
37. Метод опорных векторов. Разделяющая гиперплоскость. Зазор и опорные векторы. Каноническая гиперплоскость.
38. Метод опорных векторов. Линейный и разделимый случай. Классификатор метода опорных векторов.
39. Метод опорных векторов с мягким зазором. Прямая задача оптимизации. Оптимизация методом Ньютона.
40. Метод опорных векторов с мягким зазором. Двойственная задача оптимизации. Градиентный подъем и стохастический градиентный подъем.

Критерии оценки итогового тестирования

Проверяется правильность и полнота ответов на вопросы билета.

Комплект домашних работ

Домашняя работа № 1. Расчет статистических показателей и визуализация заданного набора данных.

- Задание:
- Дано математическое ожидание $\mathbf{a}$ и корреляционная матрица $\mathbf{R}$ двумерного гауссовского распределения.
- Постройте n значений случайных признаков $\mathbf{X}$ и $\mathbf{Y}$, имеющих двумерное гауссовское распределение с математическим ожиданием $\mathbf{a}$ и корреляционной матрицей $\mathbf{R}$.
- Визуализируйте построенный набор данных на плоскости в виде набора точек с координатами $\{(\mathbf{x}_i, \mathbf{y}_i)\}, i=1, \dots, n$.
- Вычислите и выведите на экран для построенных данных математические ожидания, дисперсии, а также корреляцию между данными.
- Считайте статистические данные для двух признаков из заданного набора данных репозитория UCI. Если в записи значение какого-либо из признаков не определено (символ "?"), то следует пропустить данную запись.
- Изобразите считанные из набора данные в виде точек на плоскости.
- Вычислите и выведите на экран для двух признаков математические ожидания, дисперсии, а также корреляцию между признаками.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 2. Использование метода главных компонент для снижения размерности данных.

- Задание:
- Проанализируйте заданный набор данных из репозитория UCI. Считайте из набора данных для последующего анализа только числовые признаки.
- Если в данных значение какого-либо из признаков не определено (символ “?”), то пропустите данную запись.
- Найдите 5 признаков, имеющих наибольшую дисперсию. Если числовых признаков меньше, чем 5, то используйте все числовые признаки.
- Для набора данных, состоящего из пяти признаков с наибольшей дисперсией, найдите размерность метода главных компонент, для которой доля объясняемой дисперсии будет не менее 95%.
- Пользуясь методом главных компонент, снизьте размерность набора данных до двух признаков и изобразите полученный набор данных в виде точек на плоскости. Отображайте точки различных классов разными цветами.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 3. Поиск ассоциативных правил при помощи алгоритмов Apriori, Eclat, Declat, FPGrowth.

- Задание:
- Скачайте заданный набор данных из репозитория UCI. Считайте из набора данных для последующего анализа только категориальные признаки, исключая числовые признаки.
- Проанализируйте набор данных и оставьте в наборе 10 категориальных, исключая признаки с неопределенными значениями и принимающие одно и то же значение для всех записей набора данных.
- Преобразуйте записи набора данных в записи транзакционной базы данных (список) следующим образом:
 - в качестве первого элемента добавьте идентификатор транзакции (порядковый номер записи в наборе)
 - в качестве второго элемента добавьте список, составленный из значений признаков записи набора с учетом указанных ниже преобразований и отсортированный по возрастанию значений
 - если признак принимает логические (булевы) значения (True/False), то подставьте вместо значения True название признака, вместо значения False ничего подставлять не нужно, например, если признак имеет название `age_gt_60`, подставьте вместо этих значения `t` (True) значение `age_gt_60`
 - если признак принимает несколько вариантов значения, то подставьте вместо текущего значения конкатенацию названия признака и его текущего значения, например, если признак имеет название `ar_c` и возможные значения признака `normal`, `elevated`, `absent`, то подставьте вместо этих значений значения `ar_c_normal`, `ar_c_elevated`, `ar_c_absent`
- В качестве минимального значения поддержки примите значение, равное 1/4 количества записей в наборе данных (относительная поддержка 0.25).
- Используя алгоритм, указанный в Вашем варианте, найдите популярные наборы данных для указанного выше значения минимальной поддержки.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 4. Решение задачи кластеризации данных при помощи алгоритма К-средних и иерархической кластеризации.

- Задание:
- Для двух классов объектов даны значения количеств объектов (точек) для каждого класса (n_1 и n_2), значения векторов математических ожиданий для каждого класса (a_1 и a_2) и корреляционные матрицы для каждого класса (R_1 и R_2) для моделируемой выборки из гауссовских случайных векторов.
- Для каждого из классов постройте значения случайных признаков X и Y , имеющие двумерное гауссовское распределение с математическим ожиданием a_i и корреляционной матрицей R_i .
- Изобразите построенные данные на плоскости в виде точек с координатами $\{(x_i, y_i)\}$, $i=1, \dots, n$ и раскрасьте их разными цветами (красным и синим) для разных классов.
- Проведите кластеризацию построенных объектов с помощью алгоритма k средних для случая, когда количество кластеров равно двум.
- Изобразите на плоскости кластеризованные объекты разными цветами для точек разных кластеров и центры кластеров.
- Найдите для построенной кластеризации таблицу сопряженности (contingency table) и вычислите показатель чистоты кластеризации.
- Считайте статистические данные для двух признаков и класса из заданного набора данных репозитория UCI. Если в записи значение какого-либо из признаков или класса не определено (символ "?"), то следует пропустить данную запись.
- Изобразите считанные из набора данные в виде точек на плоскости и раскрасьте их разными цветами для разных классов.
- Проведите кластеризацию построенных объектов с помощью алгоритма k средних для случая, когда количество кластеров равно количеству классов.
- Изобразите на плоскости кластеризованные объекты разными цветами для точек разных кластеров и центры кластеров.
- Найдите для построенной кластеризации таблицу сопряженности (contingency table) и вычислите показатель чистоты кластеризации.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 5. Решение задачи классификации данных при помощи байесовской классификации и классификации по ближайшим соседям.

- Задание:
- Для двух классов объектов даны значения количеств объектов (точек) для каждого класса (n_1 и n_2), значения векторов математических ожиданий для каждого класса (a_1 и a_2) и корреляционные матрицы для каждого класса (R_1 и R_2) для моделируемой выборки из гауссовских случайных векторов.
- Для каждого из классов постройте значения случайных признаков X и Y , имеющие двумерное гауссовское распределение с математическим ожиданием a_i и корреляционной матрицей R_i . Объедините построенные данные для двух классов в единый набор.
- Случайным образом разделите полученные данные на обучающую выборку и контрольную выборку в соотношении 80% на 20%.
- Изобразите обучающую выборку в трёхмерном пространстве в виде точек с координатами $\{(x_i, y_i, z_i)\}$, $i=1, \dots, n$ и раскрасьте их разными цветами (например, красным и синим) для разных классов.

- Произведите классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма наивной байесовской классификации.
- Изобразите объекты контрольной выборки в трёхмерном пространстве разными цветами и маркерами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Считайте статистические данные для трех признаков и класса из заданного набора данных репозитория UCI. Если в какой-либо записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 75% на 25%.
- Изобразите обучающую выборку в трёхмерном пространстве в виде точек с координатами $\{(x_i, y_i, z_i)\}$, $i=1, \dots, n$ и раскрасьте их разными цветами для разных классов.
- Произведите классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма **K** ближайших соседей для **K**, равного удвоенному количеству классов в наборе, но не менее 10.
- Изобразите объекты контрольной выборки в трёхмерном пространстве разными цветами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 6. Решение задачи классификации данных при помощи обучения дерева решений.

- Задание:
- Считайте статистические данные для двух признаков и класса из заданного набора данных репозитория UCI. Если в какой-либо записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 70% на 30%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами $\{(x_i, y_i)\}$, $i=1, \dots, n$ и раскрасьте их разными цветами для разных классов.
- Обучите классификатор DecisionTreeClassifier на обучающей выборке, ограничивая глубину дерева значением 5.
- Выполните визуализацию полученного дерева решений.
- Произведите классификацию объектов контрольной выборки с помощью обученного классификатора DecisionTreeClassifier.
- Изобразите объекты контрольной выборки на плоскости разными цветами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 7. Решение задач классификации и прогнозирования данных при помощи регрессионного анализа.

- Задание:
- Даны количество объектов (точек) n , параметры a и b и дисперсия гауссовского белого шума σ^2 .
- Смоделируйте точки $\{(x_i, y_i)\}$, $i=1, \dots, n$ согласно модели $y=ax+b+\epsilon$, где в качестве ϵ используется гауссовский белый шум (нормально распределенная случайная величина) с нулевым математическим ожиданием и заданной дисперсией σ^2 . Значения x_i выбираются через равные промежутки на отрезке $[0;1]$.
- Визуализируйте на одном графике точки (x_i, y_i) и прямую $y=ax+b$ при $x \in [0, 1]$.
- Случайным образом разделите полученные данные на обучающую выборку и контрольную выборку в соотношении 75% на 25%.
- Постройте модель линейной регрессии $y=a'x+b'+\epsilon$ на обучающей выборке.
- Выведите на экран полученные значения a' , b' и сравните их с первоначальными значениями a , b .
- Визуализируйте на одном графике точки (x_i, y_i) из контрольной выборки и прямую $y = a'x+b'$ при $x \in [0, 1]$.
- Вычислите и выведите на экран показатели MSE, MAE и коэффициент детерминации.
- Считайте данные для независимой переменной (предиктора) и зависимой переменной из заданного набора данных репозитория UCI.
- Масштабируйте зависимую переменную на диапазон от 0.001 до 0.999
- Используйте для построения модели три подхода:
 - линейную регрессию
 - полиномиальную регрессию (степень полинома $\text{degree}=2$)
 - преобразование зависимой переменной при помощи логистической функции и последующего применения линейной регрессии
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 70% на 30%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами (x_i, y_i) .
- Постройте на обучающей выборке различные модели прогнозирования значений зависимой переменной.
- Визуализируйте на одном графике разными цветами точки (x_i, y_i) из контрольной выборки, а также точки с прогнозируемыми значениями зависимой переменной для трех моделей, соединенные линиями (для улучшения картинки может потребоваться сортировка точек контрольной выборки по возрастанию независимой переменной).
- Вычислите и сравните значения показателей MSE, MAE и коэффициента детерминации для различных моделей. Определите лучшую модель по показателю коэффициента детерминации.
- Подготовьте отчет о выполнении заданий домашней работы.

Домашняя работа № 8. Решение задачи классификации данных при помощи метода опорных векторов.

- Задание:
- Считайте данные (два признака и метки) из заданного набора данных репозитория UCI.

- Разбейте метки на два класса (положительный и отрицательный) и подготовьте набор данных для обучения бинарного классификатора.
- При необходимости масштабируйте значения признаков при помощи StandardScaler.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 80% на 20%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами (x_i, y_i) с использованием разных цветов для положительного и отрицательного класса.
- Обучите бинарный классификатор метода опорных векторов LinearSVC на обучающей выборке.
- Произведите классификацию объектов контрольной выборки с помощью обученного классификатора LinearSVC.
- Изобразите объекты контрольной выборки на плоскости разными цветами для положительного и отрицательного классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Для проведенной бинарной классификации постройте ROC-кривую и выведите на рисунке показатель AUC ROC (площадь под кривой).
- Подготовьте отчет о выполнении заданий домашней работы..

Критерии оценки выполнения домашних работ

Оценивается полнота выполнения работы, оформление результатов, наличие примеров использования, полнота ответов на контрольные вопросы, если это предусмотрено заданием.