

**Рабочая программа дисциплины (модуля) «Анализ данных и машинное обучение»,
включая оценочные материалы**

1. Требования к результатам обучения по дисциплине (модулю)

1.1. Перечень компетенций, формируемых дисциплиной (модулем) в процессе освоения образовательной программы

Группа компетенций	Категория компетенций	Коды и содержание компетенций
Универсальные	-	-
Общепрофессиональные	-	ОПК-7. Способен применять в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой
	-	ОПК-8. Способен осуществлять поиск, хранение, обработку и анализ информации из различных источников и баз данных, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий
Профессиональные	-	-

1.2. Компетенции и индикаторы их достижения, формируемых дисциплиной (модулем) в процессе освоения образовательной программы

Код компетенции	Код индикатора компетенции	Содержание индикатора компетенции
ОПК-7	ОПК-7.1	Анализирует практики использования основных концепций, принципов, теорий и фактов, связанных с информатикой, в профессиональной деятельности
ОПК-7	ОПК-7.2	Способен применять в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой.
ОПК-7	ОПК-7.3	Использует в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой.
ОПК-8	ОПК-8.1	Понимает методы поиска, хранения, обработки и анализа информации из различных источников и баз данных
ОПК-8	ОПК-8.2	Формирует решения задач поиска, хранения и анализа информации из различных источников и баз данных
ОПК-8	ОПК-8.3	Использует современные информационные, компьютерные и сетевые технологии для поиска, хранения и анализа информации из различных источников и баз данных

1.3. Результаты обучения по дисциплине (модулю)

Цель изучения дисциплины (модуля) – освоение обучающимися современных технологий для обработки и анализа информации и эффективных методов ее обработки с применением современных ЭВМ, а также формирование целостной системы знаний в области создания, накопления, обработки и использования информационных ресурсов.

В результате изучения дисциплины (модуля) обучающийся должен

знать:

- понятие обобщенного метрического классификатора, больших данных и их свойства; алгоритмы метрической классификации; основные принципы построения логических алгоритмов классификации; алгоритм построения дерева классификации ID 3; линейные методы классификации; особенности обеспечения эффективной работы баз и хранилищ данных; формализацию задачи машинного обучения.

уметь:

- осуществлять математическую и информационную постановку задач по обработке информации; использовать алгоритмы обработки информации для различных приложений; выбирать методы и средства для решения задач машинного обучения;

выбирать архитектуру информационной системы использующей технологии машинного обучения, осуществлять разработку алгоритмов, баз и хранилищ данных для решения задачи машинного обучения; выполнять постановку задачи машинного обучения; выбирать методы и средства для решения задач машинного обучения;

владеть:

- инструментальными средствами анализа данных; приемами программирования на языках анализа данных; методами интеллектуального анализа данных.

2. Объем, структура и содержание дисциплины (модуля)

2.1. Объем дисциплины (модуля)

<i>Виды учебной работы</i>	<i>Формы обучения</i>
	<i>Очная</i>
Общая трудоемкость: зачетные единицы/часы	3/108
Контактная работа:	72
Лекции	36
Лабораторные работы	0
Практические занятия, семинары	36
Промежуточная аттестация: зачет	0
Самостоятельная работа (СР)	36

2.2. Темы (разделы) дисциплины (модуля) с указанием отведенного на них количества часов по формам образовательной деятельности

Очная форма обучения

№ п/п	Наименование тем (разделов)	Виды учебной работы (в часах)						СР
		Контактная работа						
		Занятия лекционного типа		Занятия семинарского типа				
		Л	Иные	ПЗ	С	ЛР	Иные	
1.	Большие данные и машинное обучение	8	0	8	0	0	0	9
2.	Метрические методы классификации	10	0	10	0	0	0	9
3.	Логические методы классификации	8	0	8	0	0	0	9
4.	Линейные методы классификации	10	0	10	0	0	0	9

Примечания:

Л – лекции, ПЗ – практические занятия, С – семинары, ЛР – лабораторные работы, СР – самостоятельная работа.

2.3. Содержание дисциплины (модуля), структурированное по темам (разделам) и видам работ

Содержание лекционного курса

№ п/п	Наименование тем (разделов)	Содержание лекционного курса
1.	Большие данные и машинное обучение	Большие данные. Свойства больших данных. Машинное обучение, формализация задачи машинного обучения. Признаковое описание объекта. Ответы и типы задач машинного обучения. Модель алгоритмов. Метод обучения. Этап обучения и этап применения. Функционалы качества. Сведение задачи обучения к задаче оптимизации. Переобучение и обобщение. Пример переобучения (Рунге).
2.	Метрические методы классификации	Формализация задачи. Обобщенный метрический классификатор. Метод ближайшего соседа. Метод квзвешенных ближайших соседей. Метод парзеновского окна.
3.	Логические методы классификации	Логическая закономерность. Основы вопросы построения логических алгоритмов классификации. Виды закономерностей. Критерии информативности: простые критерии, статистический критерий, энтропийный критерий. Где находятся закономерности в (p, n)-плоскости. Схема

		локального поиска информативных закономерностей.
4.	Линейные методы классификации	Задача построения разделяющей поверхности. Задача построения разделяющей поверхности. Минимизация эмпирического риска. Непрерывные аппроксимации пороговой функции потерь. Линейный классификатор. Персептрон. Устройство нервной клетки.

Содержание занятий семинарского типа

№ п/п	Наименование тем (разделов)	Тип	Содержание занятий семинарского типа
1.	Большие данные и машинное обучение	ПЗ	Эмпирические оценки обобщающей способности. Примеры задач машинного обучения: задачи классификации и регрессии; задачи ранжирования. Эксперименты в машинном обучении: эксперименты на реальных и синтетических данных.
2.	Метрические методы классификации	ПЗ	Метод потенциальных функций. Отбор эталонных объектов. Понятие отступа объекта. Типы объектов в зависимости от отступа. Отбор эталонов, алгоритм STOLP. Задача выбора метрики. Жадное добавление признаков.
3.	Логические методы классификации	ПЗ	Определение бинарного решающего дерева. Жадный алгоритм построения дерева ID 3. Варианты критериев ветвления в ID 3. Обработка пропусков, алгоритм обработки пропусков на этапе обучения и этапе классификации. Алгоритм ID3: достоинства и недостатки. Стратегии редукции решающих деревьев. Небрежные решающие деревья. Бинаризация вещественного признака.
4.	Линейные методы классификации	ПЗ	Линейная модель нейрона МакКаллока-Питтса. Алгоритм StochasticGradient. Дельта-правило ADALINE. Правило Хебба. SG: инициализация весов. SG: проблемы переобучения. Принцип максимума правдоподобия. Оптимальная разделяющая гиперплоскость. Метод SVM. Нелинейное обобщение SVM.

Содержание самостоятельной работы

№ п/п	Наименование тем (разделов)	Содержание самостоятельной работы
1.	Большие данные и машинное обучение	Эксперименты в машинном обучении: эксперименты на реальных и синтетических данных.
2.	Метрические методы классификации	Задача выбора метрики. Жадное добавление признаков.
3.	Логические методы классификации	Стратегии редукции решающих деревьев. Небрежные решающие деревья. Бинаризация вещественного признака.
4.	Линейные методы классификации	Оптимальная разделяющая гиперплоскость. Метод SVM. Нелинейное обобщение SVM.

3. Оценочные материалы для проведения текущего контроля успеваемости и промежуточной аттестации обучающихся по дисциплине (модулю)

По дисциплине (модулю) предусмотрены следующие виды контроля качества освоения:

- текущий контроль успеваемости;
- промежуточная аттестация обучающихся по дисциплине (модулю).

3.1. Оценочные материалы для проведения текущей аттестации по дисциплине (модулю)

№ п/п	Контролируемые темы (разделы)	Наименование оценочного средства
1.	Большие данные и машинное обучение	Устный опрос. Кейсы. Дискуссионные процедуры
2.	Метрические методы классификации	Устный опрос. Кейсы. Дискуссионные процедуры
3.	Логические методы классификации	Устный опрос. Кейсы. Дискуссионные процедуры
4.	Линейные методы классификации	Устный опрос. Кейсы. Дискуссионные процедуры

3.1.1 Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности в процессе текущего контроля успеваемости

Устный опрос. Кейсы (ситуации и задачи с заданными условиями). Дискуссионные процедуры (круглый стол, дискуссия, полемика, диспут, дебаты, мини-конференции) Предобработка данных в Pandas.

Вопросы для устного опроса:

1. Использование Phyton для анализа данных.
2. Основные библиотеки Phyton: Scikit-learn, NumPy, SciPy, matplotlib, pandas.
3. Дистрибутив Anaconda.
4. Pandas: базовые методы.
5. Pandas: индексация и извлечение данных.
6. Pandas: применение функций к ячейкам, столбцам и строкам.
7. Pandas: группировка данных.
8. Pandas: таблицы сопряженности.
9. Pandas: сводные таблицы.
10. DataFrame в Pandas.

Вопросы для групповой дискуссии:

1. Почему язык Phyton?
2. Для чего используется библиотека NumPy?
3. Для чего используются библиотеки SciPy, matplotlib, pandas?
4. Как выполняется индексация и извлечение данных в Pandas?
5. Как выполняется группировка данных в Pandas?
6. Для чего применяются таблицы сопряженности в Pandas?
7. Как выполняется загрузка данных в DataFrame библиотеке Pandas?
8. Как выполняется доступ к столбцам DataFrame библиотеке Pandas?

Кейс-задание

1. Анализ данных по доходу населения UCI Adult. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Adult. Список вопросов:
 - Каков средний возраст (признак age) женщин?
 - Какова доля граждан Германии (признак native-country)?
 - Постройте гистограмму распределения (bar plot) образования людей (признак education).
 - Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50К в год (признак salary) и тех, кто получает менее 50К в год?
 - Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)
 - Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.
 - Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.
 - Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?
 - Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).
2. Анализ данных по пассажирам Титаника. В задании предлагается с помощью

Pandas ответить на несколько вопросов по данным репозитория UCI Titanic. Список вопросов:

- Какое количество мужчин и женщин ехало на корабле?
- Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров.
- Какую долю пассажиры первого класса составляли среди всех пассажиров?
- Какого возраста были пассажиры?
- Посчитайте среднее и медиану возраста пассажиров.
- Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.
- Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name).

Метрические методы классификации в Scikit-learn.

Вопросы для устного опроса:

1. Метрические методы классификации.
2. Признаковые описания объекта.
3. Гипотеза компактности.
4. Метрики, виды метрик.
5. Весовая Евклидова метрика, метрика Минковского.
6. Масштабирование признаков.
7. Метод k ближайших соседей.
8. Реализация kNN в классе `sklearn.neighbors.KNeighborsClassifier`.
9. Кросс-валидация.
10. Алгоритм выполнения кросс-валидации по блокам.
11. Вычисление ошибки на разбиениях.

Вопросы для групповой дискуссии:

1. В чем идея гипотезы компактности?
2. В чем состоит смысл обучения в метрических методах?
3. Для чего используется библиотека Scikit-learn?
4. Для чего выполняют масштабирование признаков?
5. Как обычно выполняется масштабирование количественных признаков?
6. В каком классе Scikit-learn реализован метод kNN?
7. Какой параметр метода k ближайших соседей, задает число соседей для построения прогноза?
8. В чем смысл кросс-валидации?
9. Как вычисляется весовая Евклидова метрика?
10. Приведите формулу Метрики Минковского. Что является ее параметром?

Кейс-задание

Задание 1.

1. В этом задании нужно подобрать оптимальное значение k для алгоритма kNN. Будем использовать набор данных Wine, где требуется предсказать сорт винограда, из которого изготовлено вино, используя результаты химических анализов.
2. Выполните следующие шаги:
 - Загрузите выборку Wine по адресу <https://archive.ics.uci.edu/ml/machinelearning-databases/wine/wine.data>
 - Извлеките из данных признаки и классы. Класс записан в первом столбце (три варианта), признаки — в столбцах со второго по последний. Более подробно о сути признаков можно прочитать по адресу <https://archive.ics.uci.edu/ml/datasets/Wine>
 - Оценку качества необходимо провести методом кроссвалидации по 5 блокам (5-fold). Создайте генератор разбиений, который перемешивает выборку перед формированием блоков (`shuffle=True`). Для воспроизводимости результата, создавайте генератор KFold с фиксированным параметром `random_state=42`. В

качестве меры качества используйте долю верных ответов (accuracy).

- Найдите точность классификации на кросс-валидации для метода k ближайших соседей (`sklearn.neighbors.KNeighborsClassifier`), при k от 1 до 50. При каком k получилось оптимальное качество? Чему оно равно (число в интервале от 0 до 1)?
- Произведите масштабирование признаков с помощью функции `sklearn.preprocessing.scale`. Снова найдите оптимальное k на кросс-валидации.
- Какое значение k получилось оптимальным после приведения признаков к одному масштабу? Как изменилось значение качества? Приведите ответы на вопросы.

Задание 2.

1. Нам понадобится решать задачу регрессии с помощью метода k ближайших соседей – воспользуемся для этого классом `sklearn.neighbors.KNeighborsRegressor`.
2. Метрика задается с помощью параметра `metric`, нас будет интересовать значение `'minkowski'`. Параметр метрики Минковского задается с помощью параметра `p` данного класса.
3. Инструкция по выполнению
 - Мы будем использовать в данном задании набор данных Boston, где нужно предсказать стоимость жилья на основе различных характеристик расположения (загрязненность воздуха, близость к дорогам и т.д.). Подробнее о признаках можно почитать по адресу <https://archive.ics.uci.edu/ml/datasets/Housing>
 - Загрузите выборку Boston с помощью функции `sklearn.datasets.load_boston()`. Результатом вызова данной функции является объект, у которого признаки записаны в поле `data`, а целевой вектор — в поле `target`.
 - Приведите признаки в выборке к одному масштабу при помощи функции `sklearn.preprocessing.scale`.
 - Переберите разные варианты параметра метрики `p` по сетке от 1 до 10 с таким шагом, чтобы всего было протестировано 200 вариантов (используйте функцию `numpy.linspace`). Используйте `KNeighborsRegressor` с `n_neighbors=5` и `weights='distance'` - данный параметр добавляет в алгоритм веса, зависящие от расстояния до ближайших соседей. В качестве метрики качества используйте среднеквадратичную ошибку (параметр `scoring='mean_squared_error'` у `cross_val_score`; при использовании библиотеки `scikit-learn` версии 18.0.1 и выше необходимо указывать `scoring='neg_mean_squared_error'`). Качество оценивайте, как и в предыдущем задании, с помощью кросс-валидации по 5 блокам с `random_state = 42`, не забудьте включить перемешивание выборки (`shuffle=True`).
 - Определите, при каком `p` качество на кросс-валидации оказалось оптимальным. Обратите внимание, что `cross_val_score` возвращает массив показателей качества по блокам; необходимо сделать массив показателей качества по блокам; необходимо максимизировать среднее этих показателей.

Деревья решений. Важность признаков

Вопросы для устного опроса:

1. Логическая закономерность.
2. Основы построения логических алгоритмов классификации.
3. Определение бинарного решающего дерева.
4. Реализация решающих деревьев в библиотеке `scikit-learn`.
5. Важность признаков.
6. Пропуски в данных.

Вопросы для групповой дискуссии:

1. В каких классах `scikit-learn` реализуются решающие деревья для задач классификации и регрессии?
2. С помощью какой функции `scikit-learn` реализуется обучение модели решающих деревьев?
3. Какая переменная содержит массив "важностей" признаков?

4. С помощью какой функции можно проверить, является ли число `nan`?
5. Основы вопросы построения логических алгоритмов классификации.
6. Определение бинарного решающего дерева.
7. Реализация решающих деревьев в библиотеке `scikit-learn`.
8. Важность признаков.

Кейс-задание

1. Загрузите выборку из файла `titanic.csv` с помощью пакета `Pandas`.
2. Оставьте в выборке четыре признака: класс пассажира (`Pclass`), цену билета (`Fare`), возраст пассажира (`Age`) и его пол (`Sex`).
3. Обратите внимание, что признак `Sex` имеет строковые значения.
4. Выделите целевую переменную — она записана в столбце `Survived`.
5. В данных есть пропущенные значения — например, для некоторых пассажиров неизвестен их возраст. Такие записи при чтении их в `pandas` принимают значение `nan`. Найдите все объекты, у которых есть пропущенные признаки, и удалите их из выборки.
6. Обучите решающее дерево с параметром `random_state=241` и остальными параметрами по умолчанию.
7. Вычислите важности признаков и найдите два признака с наибольшей важностью. Их названия будут ответами для данной задачи (в качестве ответа укажите названия признаков через запятую без пробелов).

Линейная классификация. Нормализация признаков.

Вопросы для устного опроса:

1. Линейные алгоритмы классификации.
2. Персептрон.
3. Нормализация признаков. Стандартизация признаков.
4. Реализация линейных классификаторов в библиотеке `scikit-learn`.
5. Метрика качества.
6. Метод опорных векторов.
7. Опорные объекты.

Вопросы для групповой дискуссии:

1. Сформулируйте постановку задачи линейной классификации.
2. Как выполняется стандартизация признаков?
3. В каком классе `scikit-learn` реализуется персептрон?
4. Для чего используется функция `sklearn.metrics.accuracy_score`?
5. Каким классом удобно воспользоваться для стандартизации признаков?
6. На что направлен функционал, который он оптимизирует метод опорных векторов?
6. Какие объекты называют опорными?

Кейс-задание

Задание 1.

1. Загрузите обучающую и тестовую выборки из файлов `perceptrontrain.csv` и `perceptron-test.csv`. Целевая переменная записана в первом столбце, признаки — во втором и третьем.
2. Обучите персептрон со стандартными параметрами и `random_state=241`.
3. Подсчитайте качество (долю правильно классифицированных объектов, `accuracy`) полученного классификатора на тестовой выборке.
4. Нормализуйте обучающую и тестовую выборку с помощью класса `StandardScaler`.
5. Обучите персептрон на новых выборках. Найдите долю правильных ответов на тестовой выборке.
6. Найдите разность между качеством на тестовой выборке после нормализации и качеством до нее.

Задание 2.

1. Загрузите выборку из файла `svm-data.csv`. В нем записана двумерная выборка

- (целевая переменная указана в первом столбце, признаки — во втором и третьем).
2. Обучите классификатор линейным ядром, параметром $C=100000$ и `random_state=241`. Такое значение параметра нужно использовать, чтобы убедиться, что SVM работает с выборкой как с линейно разделимой. При более низких значениях параметра алгоритм будет настраиваться с учетом слагаемого в функционале, штрафующего за маленькие отступы, из-за чего результат может не совпасть с решением классической задачи SVM для линейно разделимой выборки.
 3. Найдите номера объектов, которые являются опорными (нумерация с единицы).

3.1.2. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности в ходе текущего контроля успеваемости

Устный ответ

Оценка знаний предполагает дифференцированный подход к обучающемуся, учет его индивидуальных способностей, степень усвоения и систематизации основных понятий и категорий по дисциплине. Кроме того, оценивается не только глубина знаний поставленных вопросов, но и умение использовать в ответе практический материал. Оценивается культура речи, владение навыками ораторского искусства.

Критерии оценивания: последовательность, полнота, логичность изложения, анализ различных точек зрения, самостоятельное обобщение материала, использование профессиональных терминов, культура речи, навыки ораторского искусства. Изложение материала без фактических ошибок.

Оценка «отлично» ставится в случае, когда материал излагается исчерпывающе, последовательно, грамотно и логически стройно, при этом раскрываются не только основные понятия, но и анализируются точки зрения различных авторов. Обучающийся не затрудняется с ответом, соблюдает культуру речи.

Оценка «хорошо» ставится, если обучающийся твердо знает материал, грамотно и по существу излагает его, знает практическую базу, но при ответе на вопрос допускает несущественные погрешности.

Оценка «удовлетворительно» ставится, если обучающийся освоил только основной материал, но не знает отдельных деталей, допускает неточности, недостаточно правильные формулировки, нарушает последовательность в изложении материала, затрудняется с ответами, показывает отсутствие должной связи между анализом, аргументацией и выводами.

Оценка «неудовлетворительно» ставится, если обучающийся не отвечает на поставленные вопросы.

Кейсы (ситуации и задачи с заданными условиями)

Обучающийся должен уметь выделить основные положения из текста задачи, которые требуют анализа и служат условиями решения. Исходя из поставленного вопроса в задаче, попытаться максимально точно определить проблему и соответственно решить ее.

Задачи могут решаться устно и/или письменно. При решении задач также важно правильно сформулировать и записать вопросы, начиная с более общих и, кончая частными.

Критерии оценивания – оценка учитывает методы и средства, использованные при решении ситуационной, проблемной задачи.

Оценка «отлично» ставится в случае, когда обучающийся выполнил задание (решил задачу), используя в полном объеме теоретические знания и практические навыки, полученные в процессе обучения.

Оценка «хорошо» ставится, если обучающийся в целом выполнил все требования, но не совсем четко определяется опора на теоретические положения, изложенные в научной литературе по данному вопросу.

Оценка «удовлетворительно» ставится, если обучающийся показал положительные результаты в процессе решения задачи.

Оценка «неудовлетворительно» ставится, если обучающийся не выполнил все требования.

Дискуссионные процедуры

Круглый стол, дискуссия, полемика, диспут, дебаты, мини-конференции являются средствами, позволяющими включить обучающихся в процесс обсуждения спорного вопроса, проблемы и оценить их умение аргументировать собственную точку зрения. Задание дается заранее, определяется круг вопросов для обсуждения, группы участников этого обсуждения.

Дискуссионные процедуры могут быть использованы для того, чтобы студенты:

– лучше поняли усвояемый материал на фоне разнообразных позиций и мнений, не обязательно достигая общего мнения;

– смогли постичь смысл изучаемого материала, который иногда чувствуют интуитивно, но не могут высказать вербально, четко и ясно, или конструировать новый смысл, новую позицию;

– смогли согласовать свою позицию или действия относительно обсуждаемой проблемы.

Критерии оценивания – оцениваются действия всех участников группы. Понимание проблемы, высказывания и действия полностью соответствуют заданным целям. Соответствие реальной действительности решений, выработанных в ходе игры. Владение терминологией, демонстрация владения учебным материалом по теме игры, владение методами аргументации, умение работать в группе (умение слушать, конструктивно вести беседу, убеждать, управлять временем, бесконфликтно общаться), достижение игровых целей, (соответствие роли – при ролевой игре). Ясность и стиль изложения.

Оценка «отлично» ставится в случае, когда все требования выполнены в полном объеме.

Оценка «хорошо» ставится, если обучающиеся в целом демонстрируют понимание проблемы, высказывания и действия полностью соответствуют заданным целям. Решения, выработанные в ходе игры, полностью соответствуют реальной действительности. Но некоторые объяснения не совсем аргументированы, нарушены нормы общения, нарушены временные рамки, нарушен стиль изложения.

Оценка «удовлетворительно» ставится, если обучающиеся в целом демонстрируют понимание проблемы, высказывания и действия в целом соответствуют заданным целям. Однако, решения, выработанные в ходе игры, не совсем соответствуют реальной действительности. Некоторые объяснения не совсем аргументированы, нарушены временные рамки, нарушен стиль изложения.

Оценка «неудовлетворительно» ставится, если обучающиеся не понимают проблему, их высказывания не соответствуют заданным целям.

3.2. Оценочные материалы для проведения промежуточной аттестации

3.2.1. Критерии оценки результатов обучения по дисциплине (модулю)

Шкала оценивания	Результаты обучения	Показатели оценивания результатов обучения
ОТЛИЧНО	Знает:	- обучающийся глубоко и всесторонне усвоил материал, уверенно, логично, последовательно и грамотно его излагает, опираясь на знания основной и дополнительной литературы, - на основе системных научных знаний делает квалифицированные выводы и обобщения, свободно оперирует категориями и понятиями.
	Умеет:	- обучающийся умеет самостоятельно и правильно решать учебно-профессиональные задачи или задания, уверенно, логично, последовательно и аргументировано излагать свое решение, используя научные понятия, ссылаясь на нормативную базу.
	Владеет:	- обучающийся владеет рациональными методами (с использованием рациональных методик) решения сложных профессиональных задач,

		<p>представленных деловыми играми, кейсами и т.д.;</p> <p>При решении продемонстрировал навыки</p> <ul style="list-style-type: none"> - выделения главного, - связкой теоретических положений с требованиями руководящих документов, - изложения мыслей в логической последовательности, - самостоятельного анализа факты, событий, явлений, процессов в их взаимосвязи и диалектическом развитии.
ХОРОШО	Знает:	<ul style="list-style-type: none"> - обучающийся твердо усвоил материал, достаточно грамотно его излагает, опираясь на знания основной и дополнительной литературы, - затрудняется в формулировании квалифицированных выводов и обобщений, оперирует категориями и понятиями, но не всегда правильно их верифицирует.
	Умеет:	<ul style="list-style-type: none"> - обучающийся умеет самостоятельно и в основном правильно решать учебно-профессиональные задачи или задания, уверенно, логично, последовательно и аргументировано излагать свое решение, не в полной мере используя научные понятия и ссылки на нормативную базу.
	Владеет:	<ul style="list-style-type: none"> - обучающийся в целом владеет рациональными методами решения сложных профессиональных задач, представленных деловыми играми, кейсами и т.д.; <p>При решении смог продемонстрировать достаточность, но не глубинность навыков,</p> <ul style="list-style-type: none"> - выделения главного, - изложения мыслей в логической последовательности, - связки теоретических положений с требованиями руководящих документов, - самостоятельного анализа факты, событий, явлений, процессов в их взаимосвязи и диалектическом развитии.
УДОВЛЕТВОРИТЕЛЬНО	Знает:	<ul style="list-style-type: none"> - обучающийся ориентируется в материале, однако затрудняется в его изложении; - показывает недостаточность знаний основной и дополнительной литературы; - слабо аргументирует научные положения; - практически не способен сформулировать выводы и обобщения; - частично владеет системой понятий.
	Умеет:	<ul style="list-style-type: none"> - обучающийся в основном умеет решить учебно-профессиональную задачу или задание, но допускает ошибки, слабо аргументирует свое решение, недостаточно использует научные понятия и руководящие документы.
	Владеет:	<ul style="list-style-type: none"> - обучающийся владеет некоторыми рациональными методами решения сложных профессиональных задач, представленных деловыми играми, кейсами и т.д.; <p>При решении продемонстрировал недостаточность навыков</p> <ul style="list-style-type: none"> - выделения главного, - изложения мыслей в логической последовательности, - связки теоретических положений с требованиями руководящих документов, - самостоятельного анализа факты, событий, явлений, процессов в их взаимосвязи и диалектическом развитии.
НЕУДОВЛЕТВОРИТЕЛЬНО	Знает:	<ul style="list-style-type: none"> - обучающийся не усвоил значительной части материала; - не может аргументировать научные положения; - не формулирует квалифицированных выводов и обобщений; - не владеет системой понятий.
	Умеет:	<p>обучающийся не показал умение решать учебно-профессиональную задачу или задание.</p>
	Владеет:	<p>не выполнены требования, предъявляемые к навыкам, оцениваемым «удовлетворительно».</p>

3.2.2. Контрольные задания и/или иные материалы для проведения промежуточной аттестации

Список вопросов для устных ответов (варианты теста)

1. Основные понятия – информация, данные, знания. Виды информации. Обработка данных и ее виды. Data Mining. Классификация задач Data Mining.
2. Модели процессов обработки данных. Модель: конечные автоматы.
3. Модели процессов обработки данных. Модель: сети Петри.
4. Задачи обработки данных различных типов. Прикладные области обработки данных. Оцифровка сигналов. Теорема Котельникова.
5. Базы данных. OLTP – системы. Неэффективность OLTP для анализа данных. Определение и свойства хранилищ данных.
6. Физические и виртуальные хранилища данных (ХД). Основные проблемы создания ХД.
7. Витрины данных.
8. Данные в хранилищах данных. ETL процесс.
9. Представление данных в виде гиперкуба. Операции над гиперкубом. Пример. Технология OLAP. Тест FASMI.
10. Многомерное представление данных и многомерный куб. Представление данных в виде гиперкуба. Пример.
11. Основные понятия гиперкубов (OLAP кубов). Структура OLAP куба. Операции над гиперкубом.
12. Архитектура OLAP. Компоненты OLAP. MOLAP, ROLAP, HOLAP.
13. Задача анализа текстов. Этапы анализа. Предобработка текста.
14. Извлечение ключевых понятий из текста.
15. Классификация текстовых документов. Методы классификации текстовых документов.
16. Большие данные. Свойства больших данных.
17. Машинное обучение, формализация задачи машинного обучения.
18. Признаковое описание объекта. Ответы и типы задач машинного обучения. Модель алгоритмов. Метод обучения. Этап обучения и этап применения.
19. Функционалы качества. Сведение задачи обучения к задаче оптимизации.
20. Переобучение и обобщение. Пример переобучения (Рунге). Эмпирические оценки обобщающей способности.
21. Примеры задач машинного обучения: задачи классификации.
22. Примеры задач машинного обучения: задачи регрессии.
23. Примеры задач машинного обучения: задача ранжирования.
24. Эксперименты в машинном обучении: эксперименты на реальных и синтетических данных.
25. Формализация метрической классификации. Обобщенный метрический классификатор.
26. Метод ближайшего соседа.
27. Метод взвешенных ближайших соседей.
28. Метод парзеновского окна.
29. Метод потенциальных функций.
30. Отбор эталонных объектов. Понятие отступа объекта. Типы объектов в зависимости от отступа.
31. Отбор эталонов, алгоритм STOLP.
32. Логическая закономерность. Основы вопросы построения логических алгоритмов классификации. Виды закономерностей.
33. Критерии информативности: простые критерии, статистический критерий, энтропийный критерий. Схема локального поиска информативных закономерностей.
34. Определение бинарного решающего дерева. Жадный алгоритм построения дерева ID 3.
35. Варианты критериев ветвления в ID 3.

36. Алгоритм ID3: достоинства и недостатки.
37. Обработка пропусков в ID 3, алгоритм обработки пропусков на этапе обучения и этапе классификации.
38. Стратегии редукции решающих деревьев.
39. Небрежные решающие деревья.
40. Бинаризация вещественного признака.

Тексты проблемно-аналитических и (или) практических учебно-профессиональных задач

1. Используйте библиотеку Pandas выполните задания по предварительной обработке данных.
2. Анализ данных по доходу населения UCI Adult. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Adult. Список вопросов:
 - Каков средний возраст (признак age) женщин?
 - Какова доля граждан Германии (признак native-country)?
 - Постройте гистограмму распределения (bar plot) образования людей (признак education).
 - Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50К в год (признак salary) и тех, кто получает менее 50К в год?
 - Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)
 - Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.
 - Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.
 - Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?
 - Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).
3. Анализ данных по пассажирам Титаника. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Titanic. Список вопросов:
 - Какое количество мужчин и женщин ехало на корабле?
 - Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров.
 - Какую долю пассажиры первого класса составляли среди всех пассажиров?
 - Какого возраста были пассажиры?
 - Посчитайте среднее и медиану возраста пассажиров.
 - Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.
 - Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонок Name) его личное имя (First Name).

3.2.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков в ходе промежуточной аттестации

Процедура оценивания знаний (устный ответ)

Предел длительности	10 минут
---------------------	----------

Предлагаемое количество заданий	2 вопроса
Последовательность выборки вопросов из каждого раздела	Случайная
Критерии оценки	<ul style="list-style-type: none"> - требуемый объем и структура - изложение материала без фактических ошибок - логика изложения - использование соответствующей терминологии - стиль речи и культура речи - подбор примеров их научной литературы и практики
«5» если	требования к ответу выполнены в полном объеме
«4» если	в целом выполнены требования к ответу, однако есть небольшие неточности в изложении некоторых вопросов
«3» если	требования выполнены частично – не выдержан объем, есть фактические ошибки, нарушена логика изложения, недостаточно используется соответствующая терминологии

Процедура оценивания умений и навыков (решение проблемно-аналитических и практических учебно-профессиональных задач)

Предлагаемое количество заданий	1
Последовательность выборки	Случайная
Критерии оценки:	<ul style="list-style-type: none"> - выделение и понимание проблемы - умение обобщать, сопоставлять различные точки зрения - полнота использования источников - наличие авторской позиции - соответствие ответа поставленному вопросу - использование социального опыта, материалов СМИ, статистических данных - логичность изложения - умение сделать квалифицированные выводы и обобщения с точки зрения решения профессиональных задач - умение привести пример - опора на теоретические положения - владение соответствующей терминологией
«5» если	требования к ответу выполнены в полном объеме
«4» если	в целом выполнены требования к ответу, однако есть небольшие неточности в изложении некоторых вопросов. Затрудняется в формулировании квалифицированных выводов и обобщений
«3» если	требования выполнены частично – пытается обосновать свою точку зрения, однако слабо аргументирует научные положения, практически не способен самостоятельно сформулировать выводы и обобщения, не видит связь с профессиональной деятельностью

4. Учебно-методическое и материально-техническое обеспечение дисциплины (модуля)

4.1. Электронные учебные издания

1. Замятин, А. В. Интеллектуальный анализ данных : учебное пособие / А. В. Замятин. — Томск : Издательский Дом Томского государственного университета, 2020. — 194 с. — ISBN 978-5-94621-898-6. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/116889.html>. — Режим доступа: для авторизир. пользователей
2. Пальмов, С. В. Интеллектуальный анализ данных : учебное пособие / С. В. Пальмов. — Самара : Поволжский государственный университет телекоммуникаций и информатики, 2017. — 127 с. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/75376.html>. — Режим доступа: для авторизир. пользователей

4.2. Электронные образовательные ресурсы

1. Электронная библиотечная система «ЭБС ЮРАЙТ» Biblio-online.ru (ЭБС «Юрайт») [Электронный ресурс]. – URL: <https://urait.ru/>.
2. Электронно-библиотечная система ZNANIUM [Электронный ресурс]. – URL: <https://znanium.com/>.
3. Электронная библиотечная система «Консультант студента» [Электронный ресурс]. – URL: <https://www.studentlibrary.ru/>.
4. e-Library.ru: Научная электронная библиотека [Электронный ресурс]. – URL: <http://elibrary.ru/>.
5. Научная электронная библиотека «КиберЛенинка» [Электронный ресурс]. – URL: <http://cyberleninka.ru/>.
6. Информационная система «Единое окно доступа к образовательным ресурсам» [Электронный ресурс]. – URL: <http://window.edu.ru/>.
7. Федеральный центр информационно-образовательных ресурсов [Электронный ресурс]. – URL: <http://fcior.edu.ru/>.

4.3. Современные профессиональные базы данных и информационные справочные системы

Обучающимся обеспечен доступ (удаленный доступ) к ниже следующим современным профессиональным базам данных и информационным справочным системам:

1. Словари и энциклопедии на Академике [Электронный ресурс]. – URL: <http://dic.academic.ru>.
2. Система информационно-правового обеспечения «Гарант» [Электронный ресурс]. – <http://www.garant.ru/>.
3. База данных Института философии РАН: Философские ресурсы: Текстовые ресурсы: <https://iphras.ru/page52248384.htm>.

4.4. Комплект лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства

1. Лицензионное программное обеспечение: операционная система Microsoft Windows, пакет офисных приложений Microsoft Office.
2. Свободно распространяемое программное обеспечение: свободные пакеты офисных приложений Apache Open Office, LibreOffice.
3. Программное обеспечение отечественного производства: справочно-правовая система «Гарант» (Электронный периодический справочник «Система ГАРАНТ»), образовательная платформа ЮРАЙТ (Электронная библиотечная система «ЭБС ЮРАЙТ» Biblio-online.ru (ЭБС «Юрайт»)), электронно-библиотечная система ZNANIUM, электронная библиотечная система «Консультант студента».

4.5. Оборудование и технические средства обучения

Для реализации дисциплины (модуля) используются учебные аудитории для проведения учебных занятий, которые оснащены оборудованием и техническими средствами обучения, и помещения для самостоятельной работы обучающихся, которые оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечены доступом в электронную информационно-образовательную среду РХТУ им. Д.И. Менделеева. Допускается замена оборудования его виртуальными аналогами.

Наименование учебных аудиторий для проведения учебных занятий и помещений для самостоятельной работы*	Оснащенность учебных аудиторий для проведения учебных занятий и помещений для самостоятельной работы оборудованием и техническими средствами обучения
Учебные аудитории для проведения учебных занятий	Учебная аудитория укомплектована специализированной мебелью, отвечающей всем установленным нормам и требованиям, оборудованием и техническими средствами обучения (мобильное мультимедийное оборудование).
Помещение для самостоятельной работы	Помещение оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-

	образовательную среду РХТУ им. Д.И. Менделеева и к ЭБС.
--	---

* Номер конкретной аудитории указан в приказе об аудиторном фонде, расписании учебных занятий и расписании промежуточной аттестации.