МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Югорский государственный университет»

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Искусственный интеллект для анализа данных

Направление подготовки (специальности): *21.05.06 Нефтегазовые техника и технологии*

Профиль: *Разработка и эксплуатация нефтяных и газовых месторождений*

Форма обучения
Очная

Квалификация выпускника
Горный инженер
(специалист)
2025 год набора

Виды работ	Объём занятий по семестрам, час										Итого
	1	2	3	4	5	6	7	8	9	10	
Лекции						6					6
Практические (семинарские занятия)						24					24
Самостоятельная работа						42					42
Форма контроля						Дифференцированный зачет					-
Итого:						72					72
з.е.						2					2

Ханты-Мансийск, 2024 год
(город)

Предисловие

1. Программа разработана в соответствии с требованиями Федерального закона от 27.12.2012 г. № 273-ФЗ «Об образовании в Российской Федерации», федерального государственного образовательного стандарта высшего образования (ФГОС ВО) по направлению подготовки (специальности) *21.05.06 Нефтегазовые техника и технологии* утвержденного № 27 от 11.01.2018 года.

2. Разработчик(и):

Кандидат физико-математических наук, Доцент		А. С. Шевченко
ученая степень, ученое звание (при наличии)	(подпись)	(И. О. Фамилия)

3. Согласовано:

Руководитель образовательной программы по направлению подготовки 21.05.06 Нефтегазовые техника и технологии		Т. И. Романова
	(подпись)	(И. О. Фамилия)

4. Утверждаю:

Руководитель структурного подразделения Центр образовательного инжиниринга		И. Д. Лебедева
	(подпись)	(И. О. Фамилия)

Документ подписан простой электронной подписью в электронной информационно образовательной среде Elios 2.0 ФГБОУ ВО «ЮГУ»

Идентификатор документа: 39818

Подписант

Шевченко Алеся Сергеевна

Романова Татьяна Ивановна

Лебедева Илона Дмитриевна

1 Цель освоения дисциплины

Целью освоения дисциплины является изучение основных принципов сбора, хранения и обработки больших данных с помощью библиотек Python. Студенты научатся анализировать табличные данные с помощью библиотеки Pandas, познакомятся с подходами к оптимизации вычислений с помощью библиотеки Numpy, рассмотрят возможности библиотек seaborn и matplotlib для визуализации табличных данных, а также научаться применять машинное обучение для предсказания событий, прогнозирования значений и поиска закономерностей в данных.

2 Место дисциплины в структуре ОПОП

Дисциплина относится к обязательной части блока ФТД учебного плана, модуля «Дисциплины по выбору ДВ-7 (модуль саморазвития 3)».

3 Формируемые компетенции обучающегося

Планируемые результаты освоения ОПОП (компетенции), достижение которых обеспечивает дисциплина		Планируемые результаты (соотнесенные с установленными индикаторами достижения компетенции)
код компетенции	наименование компетенции	
УК-6	*Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки и образования в течение всей жизни*	*УК-6.2 З-1:* *Имеет базовые знания в отдельной сфере, выбранной для целей саморазвития.* *УК-6.2 У-1:* *Умеет применять инструменты самооценки для выстраивания траектории саморазвития в системе непрерывного образования.* *УК-6.2 В-1:* *Имеет практический опыт получения дополнительного образования для целей саморазвития.*

4 Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 2 зачетных единицы, 72 часа.

№ п/п	Тема	Трудоемкость по видам учебной работы, час					Код компетенции	Оценочные средства
		Занятия лекционного типа	Практические занятия	Лабораторные занятия	Консультации	Самостоятельная работа		

1	Введение в анализ данных. Какие задачи решаются в анализе данных, их сходства и отличия. Стандарт CRISP-DM: решение задач анализа данных. Роли в проектах по анализу данных. Среда разработки для языка Python: Anaconda, Google colab	1	1			6	УК-6.	Тест.
2	Библиотека NumPy. Базовый функционал NumPy для аналитиков данных. Преимущества NumPy.	1	5			8	УК-6.	Тест; Электронный практикум.
3	Библиотека Pandas. Базовая аналитика. Представление одномерных данных с помощью объекта Series. Преобразование табличных и многомерных данных с помощью объекта DataFrame. Группировка и агрегирование данных. Анализ временных рядов.	1	6			8	УК-6.	Тест; Электронный практикум.
4	Разведочный анализ данных. Библиотеки seaborn и matplotlib для визуализации данных.	1	4			8	УК-6.	Тест; Электронный практикум.
5	Методы машинного обучения, применяемые для анализа данных. Задача регрессии. Задача классификации. Задача кластеризации. Библиотека машинного обучения sklearn.	2	8			12	УК-6.	Тест; Электронный практикум.
	Итого	6	24			42	–	

5 Образовательные технологии, используемые при различных видах учебной работы

№ темы	Образовательная технология
1-5	Интерактивные технологии
1-5	Дистанционные технологии

6 Методические материалы по освоению дисциплины

Электронная информационно - образовательная среда представлена личным кабинетом, расположенным поссылке https://itport.ugrasu.ru, электронной библиотечной системой https://lib.ugrasu.ru, электронным каталогом Научной библиотеки ЮГУ https://irbis.ugrasu.ru и системой дистанционного обучения.

Методические материалы для обучающихся представлены в электронном виде в системе Moodle по ссылке http://eluniver.ugrasu.ru.

Методические материалы для обучающихся из числа инвалидов и лиц с ОВЗ предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации.

6.1 Методические указания к занятиям лекционного типа

Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначить вопросы, термины, материал, который вызывает трудности, пометить и попытаться найти ответ в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать его научно-педагогическому работнику на консультации, на практическом занятии.

6.2 Методические указания к практическим занятиям

Целью практических занятий является закрепление теоретических знаний и приобретение практических умений и навыков. Методические рекомендации по каждой практической работе имеют теоретическую часть, подготовленную отдельно, или указание на источник, необходимый для подготовки к соответствующему практическому занятию, с необходимыми для выполнения работы формулами, пояснениями, таблицами и графиками; алгоритм выполнения заданий. Практические задания сочетаются с теоретическими знаниями. Проведению практического занятия как правило предшествует самостоятельная работа обучающегося.

6.3 Методические указания к самостоятельной работе

В рамках самостоятельной работы обучающийся знакомится с рабочей программой, особое внимание должно уделяться целям и задачам, структуре и содержанию дисциплины. Анализируется конспект лекций, ведется подготовка ответов к контрольным вопросам, просматривается рекомендуемая литература, используются аудио-видеозаписи по заданной теме, решаются расчетно-графические задания, задачи по алгоритму и др.

7 Оценочные материалы для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины, учебно-методическое обеспечение самостоятельной работы обучающихся.

Текущий контроль контроль успеваемости обеспечивает оценивание хода освоения дисциплин (модулей). Для осуществления процедуры текущего контроля успеваемости обучающихся НПР создаются оценочные материалы (фонды оценочных средств),

позволяющие оценить достижение запланированных результатов обучения и уровень сформированности компетенций.

Промежуточная аттестация обучающихся производится в дискретные временные интервалы НПР, обеспечивающими реализацию дисциплины в форме: дифференцированный зачет.

Учебно-методическое обеспечение самостоятельной работы обучающихся предполагает предоставление студентам методических рекомендаций по изучению дисциплины, учитывающих особенности ее построения, освоения, преподавания и представлено как электронный учебно-методический комплект документов по дисциплине, размещено в системе управления обучением «Moodle» (сайт Университета по ссылке http://eluniver.ugrasu.ru) и/или в других системах управления обучением электронной информационно-образовательной среды Университета.

Обучение и контроль обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья при необходимости осуществляется с использованием специальных методов обучения и дидактических материалов, составленных с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся (обучающегося).

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц согранченными возможностями здоровья предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации.

7.1 Технологическая карта дисциплины 6-й семестр

№ п/п	Название темы	Максимальное количество баллов
Обязательный уровень (текущая аттестация)		
1	Введение в анализ данных. Какие задачи решаются в анализе данных, их сходства и отличия. Стандарт CRISP-DM: решение задач анализа данных. Роли в проектах по анализу данных. Среда разработки для языка Python: Anaconda, Google colab	4
2	Библиотека NumPy. Базовый функционал NumPy для аналитиков данных. Преимущества NumPy.	16
3	Библиотека Pandas. Базовая аналитика. Представление одномерных данных с помощью объекта Series. Преобразование табличных и многомерных данных с помощью объекта DataFrame. Группировка и агрегирование данных. Анализ временных рядов.	16
4	Разведочный анализ данных. Библиотеки seaborn и matplotlib для визуализации данных.	14
5	Методы машинного обучения, применяемые для анализа данных. Задача регрессии. Задача классификации. Задача кластеризации. Библиотека машинного обучения sklearn.	20
		70
Обязательный уровень (промежуточная аттестация)		
6	Дифференцированный зачет	30
		30
	Итого	100
Дополнительный уровень		
7	Очное участие в конференции по тематике дисциплины	10
8	Публикация научной статьи по тематике дисциплины	5
		15

Шкала оценивания результатов по балльной системе (дифференцированный зачет):
Критерии выставления оценки при промежуточной аттестации:
Отлично с 83 по 100 баллов;
Хорошо с 68 по 82 балла;
Удовлетворительно с 50 по 67 баллов;
Неудовлетворительно с 0 по 49 баллов.

7.2 Примерные тестовые задания

1. Какой из этих языков программирования обычно используют для анализа данных?
   а) Lisp
   б) Python
   в) C#
   г) JavaScript
   д) Assembler
2. Почему при реализации проекта по анализу данных приходится возвращаться на более ранние этапы после этапа валидации?
   а) Модель может иметь высокие метрики, но не соответствовать бизнес-этике, например отклоняет запрос на кредит всем женщинам.
   б) Могли быть не учтены бизнес-нюансы на этапе постановки задачи.
   в) Модель работает точно, но слишком дорога в обслуживании.
   г) Модель работает некорректно на определенной подвыборке, например на заемщиках с низкими доходами.
   д) Модель плохо работает по бизнес-метрикам.
3. Назовите тип данных категориальных столбцов в Pandas по умолчанию
   а) int64
   б) object
   в) string
   г) float64
4. Дан следующий DataFrame:
   df = pd.DataFrame({‘a’: [1,2,3,4], ‘b’: [5,6,7,8], ‘c’: [9,10,11,12]}, index=[1,2,3,4])
   Что вернет код df.iloc[1, 2]?
   а) 6
   б) 7
   в) 9
   г) 10
5. Какая структура данных Pandas используется для одномерных данных?
   а) Series
   б) DataFrame
   в) Array
   г) List
6. Как в библиотеке NumPy можно создать массив со значениеми в диапазоне от 0 до 9?
   а) np.arrange(0,10)
   б) np.arrange(10)

в) np.arrange(0,9)

г) np.linspace(0,9, 10)

7. Для каких моделей регрессии требуется нормализация признаков?

а) Линейная регрессия.

б) Решающие деревья.

в) Метод k ближайших соседей.

г) Случайный лес.

8. Какие реальные задачи можно решать с помощью регрессии?

а) Оценка вероятности дефолта клиента по кредиту.

б) Прогнозирование стоимости аренды жилья.

в) Определение объекта на картинке.

г) Построение зависимости между географическим положением торгового центра и объемом продаж.

9. Пусть в матрице ошибок TP = 5, TN = 90, FP = 10, FN = 5. Оцените метрики классификации для такой матрицы ошибок.

10. Кластеризация — это:

а) Предсказание класса.

б) Предсказание вещественного числа.

в) Группировка данных в пространстве признаков.

7.3 Примерные задания для электронного практикума

1. Скачайте набор данных, который представляет собой статистику параметров автомобилей на вторичном рынке. Набор включает ряд категориальных и численных значений, составляющих одну запись (строку).

а) Используйте функции и методы библиотеки Pandas для загрузки и начальной работы с данными.

б) Выполните визуализацию данных с использованием библиотеки Pandas. Попробуйте построить разные виды графиков для числовых признаков – скаттерограммы, гистограммы и т. д. Для скаттерограмм попробуйте использовать категориальные данные для таких параметров графиков, как оттенок (hue), размер маркера (size), тип маркера (style). Таким образом, вы можете объединить информацию о нескольких признаках в один двумерный график.

в) Попытайтесь добавить в модель дополнительные признаки на основе имеющихся. Проверьте корреляцию новых признаков с добавленными.

г) Выполните предварительную обработку данных. Сохраните результаты разных методов предварительной обработки в разные файлы, чтобы потом была возможность протестировать различные гипотезы.

2. Набор данных представляет собой статистику параметров автомобилей на вторичном рынке. Набор включает ряд категориальных и численных значений, составляющих одну запись (строку). Каждый столбец в записи – это отдельный параметр.

Среди указанных параметров приведены целевой для задачи предсказания (регрессии) – цена автомобиля.

Используйте любую из подготовленных вами моделей линейной регрессии для предсказания цены автомобилей в наборе данных Cars. Для оценки качества модели используйте отложенную выборку и несколько метрик регрессии. Сравните результаты модели при использовании только числовых признаков и при добавлении категориальных признаков с помощью One-Hot-кодирования.

Сравните работу реализованных алгоритмов с функциями библиотеки scikit-learn:

- простая линейная регрессия через метод наименьших квадратов sklearn.linear_model.LinearRegression;
- простая линейная регрессия через градиентный спуск sklearn.linear_model.SGDRegressor;
- регрессия с регуляризацией Тихонова sklearn.linear_model.Ridge;
- регрессия с L1-регуляризацией sklearn.linear_model.Lasso;
- эластичная регуляризация sklearn.linear_model.ElasticNet.

3. Набор данных представляет собой статистику параметров автомобилей на вторичном рынке. Набор включает ряд категориальных и численных значений, составляющих одну запись (строку). Каждый столбец в записи — это отдельный параметр.

Среди указанных параметров приведены целевой для задачи классификации – тип трансмиссии.

Используйте модель логистической регрессии для предсказания типа трансмиссии автомобилей в наборе данных Cars. Для оценки качества модели используйте отложенную выборку и несколько метрик классификации. Сравните результаты модели при использовании только числовых признаков и при добавлении категориальных признаков с помощью One-Hotкодирования.

Сравните работу реализованных алгоритмов с функцией библиотеки scikit-learn – логистической регрессией sklearn.linear_ model.LogisticRegression.

4. Рассчитать параметры описательной статистики для переменных набора с использованием методов библиотеки pandas языка Python. Сделать содержательные выводы.

5. Используя ресурс kaggle: https://www.kaggle.com/, выбрать один из наборов данных. Загрузить этот набор в рабочую директорию. Считать данные, определить тип данных. Описать данные набора: какие переменные в нем присутствуют, какой тип данных у этих переменных.

7.4 Примерный список вопросов, задаваемых на диф. зачете

1. Какие задачи решаются в анализе данных, их сходства и отличия.
2. Стандарт CRISP-DM: решение задач анализа данных.
3. Роли в проектах по анализу данных.
4. Среда разработки для языка Python: Anaconda, Google colab.
5. Базовый функционал NumPy для аналитиков данных. Преимущества NumPy.
6. Библиотека Pandas. Базовая аналитика.
7. Представление одномерных данных с помощью объекта Series.

8. Преобразование табличных и многомерных данных с помощью объекта DataFrame.

9. Группировка и агрегирование данных с помощью библиотеки Pandas.

10. Анализ временных рядов с помощью библиотеки Pandas.

11. Разведочный анализ данных.

12. Библиотеки seaborn и matplotlib для визуализации данных.

13. Задачи машинного обучения.

14. Задача регрессии. Метрики, критерии качества для задач регрессии. Методы решения задачи регрессии: линейная регрессия, метод k-ближайших соседей, решающие деревья.

15. Задача классификации. Метрики, критерии качества для задач классификации. Методы решения задач классификации: логистическая регрессия, решающие деревья.

16. Задача кластеризации. Внешние и внутренние метрики качества. Методы кластеризации.

17. Библиотека машинного обучения sklearn.

18. Предварительная обработка числовых и категориальных признаков.

19. Написание собственных классов для предварительной обработки.

20. Особенности работы с Pipeline.

21. Линейная регрессия в sklearn.

22. Логистическая регрессия sklearn.

23. Метод k-ближайших соседей в sklearn.

24. Деревья решений в sklearn.

8 Материально-техническое и учебно-методическое обеспечение дисциплины

8.1 Перечень учебной литературы

Наименование печатных и (или) электронных учебных изданий, методические издания, периодические издания по всем входящим в реализуемую образовательную программу учебным предметам, курсам, дисциплинам (модулям) *в соответствии с рабочими программами дисциплин, модулей, практик*		Количество экземпляров	Обеспеченность студентов учебной литературой (экземпляров на одного студента)
Электронные учебные издания, имеющиеся в электронном каталоге электронно-библиотечной системы	Алексеев, Д. С. Технологии интеллектуального анализа данных : учебник для вузов / Д. С. Алексеев, О. В. Щекочихин. - 2-е изд., стер. - Санкт-Петербург : Лань, 2024. - 176 с. - </A></A> УДК 004 ББК 32.81я73 Кл.слова (ненормированные): технологии интеллектуального анализа данных кластеризация данных искусственные нейронные сети генетические алгоритмы нечеткое моделирование scilab.	1	1
	Макшанов, А. В. Технологии интеллектуального анализа данных : учебное пособие / А. В. Макшанов, А. Е. Журавлев. - 2-е изд., стер. - Санкт-Петербург : Лань, 2022. - 212 с. - </A></A> УДК 4 ББК 32.81 Кл.слова (ненормированные): информационные технологии анализ данных matlab среда обработки алгоритм системы и модели средства реализации технологий.	1	1
	Колмогорова, С. С. Обработка данных алгоритмами искусственного интеллекта в системе интернета вещей : учебное пособие для вузов / С. С. Колмогорова. - Санкт-Петербург : Лань, 2023. - 104 с. - </A></A> УДК 004.8 ББК 32.813я73 Кл.слова	1	1

	(ненормированные): машинное обучение интернет вещей искусственный интеллект распознавание jupyter notebook.		

8.2 Современные профессиональные базы данных, информационные справочные и электронно-библиотечные системы

№	Ссылка на информационный ресурс	Наименование ресурса в электронной форме	Доступность
Электронно-библиотечные системы			
1	https://dlib.eastview.com	База данных «Ивис»	Авторизованный доступ
2	http://elibrary.ru	Научная электронная библиотека eLIBRARY.RU	Авторизованный доступ
3	https://urait.ru	Образовательная платформа Юрайт	Авторизованный доступ
4	http://www.iprbookshop.ru	ЭБС IPR SMART	Авторизованный доступ
5	http://znanium.com	ЭБС «Znanium»	Авторизованный доступ
6	https://e.lanbook.com	ЭБС «Лань»	Авторизованный доступ
Информационные справочные системы			
7	http://www.consultant.ru/	СПС КонсультантПлюс	Авторизованный доступ
Профессиональные базы данных			
8	http://garant.ugrasu.ru/	СПС Гарант	Авторизованный доступ

8.3 Перечень лицензионного и свободно распространяемого программного обеспечения, используемого при осуществлении образовательного процесса по дисциплине, в том числе отечественного производства

Python;
Visual Studio Code;
Антиплагиат.ВУЗ;

8.4 Материально-техническое обеспечение дисциплины

8.4.1 Учебная аудитория лекционного типа
компьютер/ноутбук, проектор, экран, учебная мебель, учебная доска

8.4.2 Учебная аудитория для самостоятельной работы
учебная мебель, компьютеры с выходом в интернет и доступом к электронной информационно-образовательной среде

8.4.3 Учебная аудитория для проведения практических занятий (компьютерный класс)

Учебная мебель, учебная доска, компьютеры с доступом в Интернет