Приложение 4

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный университет
им. Н.И. Лобачевского»

Факультет социальных наук

УТВЕРЖДЕНО

решением президиума Ученого совета ННГУ

протокол № 1 от 16.01.2024 г.

Рабочая программа дисциплины

Введение в машинное обучение с использованием больших данных

Уровень высшего образования
Бакалавриат

Направление подготовки / специальность
39.03.01 - Социология

Направленность образовательной программы
Социальная теория и комплексный анализ данных

Форма обучения
очная

г. Нижний Новгород

2024 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.08.02 Введение в машинное обучение с использованием больших данных относится к части, формируемой участниками образовательных отношений образовательной программы.

2. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	**Планируемые результаты обучения по дисциплине (модулю), в соответствии с индикатором достижения компетенции**		**Наименование оценочного средства**	
	Индикатор достижения компетенции (код, содержание индикатора)	**Результаты обучения по дисциплине**	**Для текущего контроля успеваемости**	**Для промежуточной аттестации**
ПК-7: Способен проводить социологический мониторинг социальной среды и разрабатывать управленческие решения	*ПК-7.1: Формирует информационные базы данных для социологического мониторинга социальной сферы* *ПК-7.2: Использует адекватные поставленным организационноуправленческим задачам методы социологического анализа* *ПК-7.3: Формулирует предложения по совершенствованию социальных процессов и отношений на основе анализа и обобщения результатов социологических исследований*	*ПК-7.1:* *Знать методики качественного и количественного анализа в социальных науках* *Уметь оценивать потребность в данных* *Владеть навыками проведения анализа данных социологического мониторинга и визуализации результатов с применением алгоритмов машинного обучения* *ПК-7.2:* *Знать основные типы данных для построения моделей* *Владеть навыками постановки задач для реализации социологического анализа с применением алгоритмов машинного обучения* *ПК-7.3:* *Владеть навыками формирования выводов на основе социологического анализа с применением алгоритмов машинного обучения*	*Практическое задание*	*Зачёт:* *Проект*

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	**2**
Часов по учебному плану	**72**
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	**16**
- занятия семинарского типа (практические занятия / лабораторные работы)	**32**
- КСР	**1**
самостоятельная работа	**23**
Промежуточная аттестация	**0** **Зачёт**

3.2. Содержание дисциплины

(структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий)

Наименование разделов и тем дисциплины	Всего (часы)	в том числе			
		Контактная работа (работа во взаимодействии с преподавателем), часы из них			Самостоятельная работа обучающегося, часы
		Занятия лекционного типа	Занятия семинарского типа (практические занятия/лабораторные работы), часы	Всего	
	ооо	ооо	ооо	ооо	ооо
Тема 1. Понятие машинного обучения: общая характеристика, задачи, виды	9	2	2	4	5
Тема 2. Машинное обучение с учителем: задачи классификации и регрессии	17	4	8	12	5
Тема 3. Машинное обучение без учителя: понижение размерности и задачи кластеризации	15	4	8	12	3
Тема. 4. Анализ и кластеризация текстов	19	4	8	12	7
Тема 5. Введение в нейронные сети	11	2	6	8	3
Аттестация	0				
КСР	1			1	
Итого	72	16	32	49	23

Содержание разделов и тем дисциплины

Тема 1. Понятие машинного обучения: общая характеристика, задачи, виды Что такое модели с учителем. Типы данных для построения регрессионных моделей. Примеры предобработки и визуализации данных для построения регрессии

Библиотеки для построения регрессионных моделей. Построение регрессионных моделей и оценка их валидности. Метрики качества моделей. Визуализация результатов построения регрессионных моделей Регуляризация регрессионных моделей – назначение и примеры применения
Тема 2. Машинное обучение с учителем: задачи классификации и регрессии Типы данных для построения логистической регрессии. Примеры предобработки и визуализации количественных и качественных переменных.
Библиотека для построения логистической регрессии. Балансировка классов с применением библиотеки SMOTE. Построение модели логистической регрессии.
Оценка качества модели логистической регрессии. Оценка качества классификатора. Roc -кривые Построение модели с применением алгоритма случайный лес – используемые библиотеки, трактовка результатов построения модели
Тема 3. Машинное обучение без учителя: понижение размерности и задачи кластеризации Задачи без учителя – общая характеристика. Виды задач без учителя. Типы данных для реализации задач. Примеры практических задач.
Задачи понижения размерности – основное назначение и геометрическая интерпретация. Задачи кластеризации – назначение, K-means++ – определение, библиотеки для реализации, реализация метода K-means++, визуализация результатов, выводы примеры прикладных задач.
Тема 4. Анализ и кластеризация текстов. Назначение анализа текстов. Понятие обработки текстов на естественном языке. Этапы обработки текстов. Кластеризация текстов -библиотеки, практическая реализация, выводы. Модели группы Word2vec
Тема 5. Введение в нейронные сети Основы обучения нейронных сетей. Архитектуры нейронных сетей. Прикладные задачи, решаемые с применением рейнонных сетей

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведенным в п. 5.

Для обеспечения самостоятельной работы обучающихся используются:
Электронные курсы, созданные в системе электронного обучения ННГУ:

Машинное обучение с использованием больших данных, https://e-learning.unn.ru/course/view.php?id=4500.

5. Фонд оценочных средств для текущего контроля успеваемости и промежуточной аттестации по дисциплине (модулю)

5.1 Типовые задания, необходимые для оценки результатов обучения при проведении текущего контроля успеваемости с указанием критериев их оценивания:

5.1.1 Типовые задания (оценочное средство - Практическое задание) для оценки сформированности компетенции ПК-7:

Задание 1. На основании данных о ценах на дома в Альбукерке (файл Albuquerque Home Prices):

- Произвести предварительную обработку данных, принять решение в отношении пропущенных значений переменных «AGE» и «TAX»
- Построить регрессионную модель, позволяющую спрогнозировать цену дома, произвести анализ модели
- Найти взаимосвязанные регрессоры, исключить незначимые переменные, построить модель для прогнозирования цены дома, сделать выводы.

Задание 2. На основании данных о ценах на бриллианты и их весе (файл diamond.dat) построить полиномиальные регрессии и выбрать наиболее удачную на основании анализа метрик качества моделей

Задание 3. Произвести кодирование качественных переменных, представленных в наборе данных о ценах на бриллианты (diamonds.csv), выполнить анализ распределения переменных, найти взаимосвязи между регрессорами средствами библиотеки Seaborn, построить регрессионные модели для прогнозирования цен на бриллианты и выбрать наиболее удачную на основе анализа метрик качества, построить график распределения ошибок в регрессионной модели, построить ридж-регрессию и регрессию Lasso, сделать вывод о том, как изменились коэффициенты в модели

Задание 4.

Выполнить процедуру регуляризации данных на сгенерированном наборе данных, найти наилучший параметр alpha, показать как меняется значение коэффициентов линейной регрессии в зависимости от параметра alpha

Задание 5. Показать как меняется средняя стандартная ошибка в зависимости от показателя степени регрессора на примере сгенерированного набора данных.

Задание 6. Показать как меняется средняя стандартная ошибка в зависимости от объема тренировочный выборки на примере сгенерированного набора данных. Построить кривую обучения

Задание 7. Выполнить стандартизацию показателей, произвести преобразования Бокса-Кокса, проверив распределение на нормальность, построить Q-Q график

(набор данных содержится в файле Churn_Modelling.xlsx).

Задание 8. Произвести классификацию данных метод K-ближайших соседей. Данные доступны по ссылке https://www.stats.govt.nz/large-datasets/csv-files-for-download/

Задание 9. Выделить главные компоненты

Для выполнения заданий 8 и 9 использовать набор данных Iris

https://gist.github.com/curran/a08a1080b88344b0c8a7

Задание 10. Используя информацию о наблюдениях за НЛО (данные содержатся по ссылке https://www.kaggle.com/datasets/NUFORC/ufo-sightings?select=scrubbed.csv) определить тип данных, построить различные типы графиков с использованием библиотек Seaborn и Matplotlib, сделать выводы.

Задание 11. Произвести проверку некоторых статистических гипотез, сделать выводы.

Наборы данных для выполнения задания представлены по ссылке **https://github.com/Laggg/jupyters-for-ds-students**

Задание 12. Построить логистическую регрессию, используя набор данных

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing. оценить качество полученной модели,

сделать выводы,

Задание 13. Произвести балансировку классов с применением библиотеки Synthetic Minority Oversampling Technique – SMOTE. Данные для выполнения задания следует найти по ссылке https://habr.com/ru/post/452392/

Задание 14. Выполнить кластеризация текста с применением пакета библиотек **NLTK** для символьной и статистической обработки естественного языка. Текст для выполнения задания содержится в файле lenta-ru-news.csv.

Задание 15. Построить классификатор с применением дерева решений на основе данных о махинациях с кредитными картами. Данные доступны по ссылке Обнаружение мошенничества с кредитными картами. Оценить качество классификаторы с применением необходимых метрик, визуализировать дерево решений, найти значимые переменные для выполнения классификации, сделать выводы.

Задание 16. Построить классификатор с применением алгоритма случайный лес на основе данных о качестве красного вина. Данные доступны по ссылке Качество красного вина. Оценить качество классификаторы с применением необходимых метрик, найти значимые переменные для выполнения классификации, сделать выводы.

Критерии оценивания (оценочное средство - Практическое задание)

Оценка	Критерии оценивания
зачтено	Практические задания выполнены без грубых ошибок
не зачтено	Практические задания выполнены с грубыми ошибками

5.2. Описание шкал оценивания результатов обучения по дисциплине при промежуточной аттестации

Шкала оценивания сформированности компетенций

Уровень сформированности компетенций (индикатора достижения компетенций)	**плохо**	**неудовлетворительно**	**удовлетворительно**	**хорошо**	**очень хорошо**	**отлично**	**превосходно**
	не зачтено		**зачтено**				
Знания	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний вследствие отказа обучающегося от ответа	Уровень знаний ниже минимальных требований. Имели место грубые ошибки	Минимально допустимый уровень знаний. Допущено много негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Ошибок нет.	Уровень знаний в объеме, превышающем программу подготовки.
Умения	Отсутствие	При решении	Продемонс	Продемонс	Продемонс	Продемонс	Продемонстр

	минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки	трированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме	трированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания в полном объеме, но некоторые с недочетами	трированы все основные умения. Решены все основные задачи. Выполнены все задания в полном объеме, но некоторые с недочетами.	трированы все основные умения. Решены все основные задачи с отдельными несущественными недочетами, выполнены все задания в полном объеме	ированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
Навыки	Отсутствие базовых навыков. Невозможность оценить наличие навыков вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки	Имеется минимальный набор навыков для решения стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач без ошибок и недочетов	Продемонстрированы навыки при решении нестандартных задач без ошибок и недочетов	Продемонстрирован творческий подход к решению нестандартных задач

Шкала оценивания при промежуточной аттестации

Оценка		**Уровень подготовки**
зачтено	**превосходно**	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно», продемонстрированы знания, умения, владения по соответствующим компетенциям на уровне выше предусмотренного программой
	отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично».
	очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо»
	хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо».
	удовлетворительно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «удовлетворительно», при этом хотя бы одна компетенция сформирована на уровне «удовлетворительно»
не зачтено	**неудовлетворительно**	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно».
	плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.3 Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения на промежуточной аттестации с указанием критериев их оценивания:

5.3.1 Типовые задания (оценочное средство - Проект) для оценки сформированности компетенции ПК-7

Самостоятельная работа обучающегося реализуется в форме выполнения проекта. Тема проекта выбирается обучающимися самостоятельно в зависимости от направления подготовки, а также решаемых научных и практических задач.

Тема проекта формулируется например

Диагностика взаимосвязи уровня экономической безопасности и отдельных показателей качества жизни населения регионов России или

Анализ отдельных показателей качества жизни населения регионов России

В результате выполнения проекта обучающийся должен продемонстрировать, что он:

- имеет навыки постановки задач машинного обучения;
- знает основные типы данных для построения моделей машинного обучения, требования к ним;
- умеет оценивать потребность в данных для реализации поставленных задач и осуществлять сбор данных.

Владеет навыками:

- проведения предобработки, анализа данных и визуализации полученных результатов с применением алгоритмов машинного обучения
- выбора моделей в зависимости от поставленных задач;
- построения моделей и оценки их качества;
- прогнозирования на основе полученной модели
- формирования выводов на основе анализа данных с применением алгоритмов машинного обучения.

Проект должен включать:

- Постановку проблемы.
- Формулировку целей и задач проекта.
- Подбор данных для анализа
- Статистический анализ и визуализацию данных.
- Обоснование выбора моделей, построение моделей, который описывали бы взаимосвязи между показателя
- Оценку качества моделей
- Прогнозирование на основе модели
- Обоснование полученных результатов
- Выводы

Проект должен в обязательно порядке содержать файлы с набором команд для построения моделей (код). Комментарии в коде обязательны.

Критерии оценивания (оценочное средство - Проект)

Оценка	Критерии оценивания
зачтено	Практические задания выполнены без грубых ошибок

Оценка	Критерии оценивания
не зачтено	Практические задания выполнены с грубыми ошибками

6. Учебно-методическое и информационное обеспечение дисциплины (модуля)

Основная литература:

1. Козлов Андрей Юрьевич. Статистический анализ данных в MS Excel : Учебник / Пензенский государственный университет; Национальный исследовательский университет "Высшая школа экономики"; Военная академия материально-технического обеспечения им. генерала армии А.В. Хрулёва, ф-л г. Пенза. - 1. - Москва : ООО "Научно-издательский центр ИНФРА-М", 2021. - 320 с. - ВО - Бакалавриат. - ISBN 978-5-16-004579-5. - ISBN 978-5-16-101024-2., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=770671&idb=0.
2. Кулаичев Алексей Павлович. Методы и средства комплексного статистического анализа данных : Учебное пособие / Московский государственный университет им. М.В. Ломоносова, биологический факультет. - 5. - Москва : ООО "Научно-издательский центр ИНФРА-М", 2022. - 484 с. - ВО - Бакалавриат. - ISBN 978-5-16-012834-4. - ISBN 978-5-16-103357-9., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=771058&idb=0.
3. Дайитбегов Дайитбег Магамедович. Компьютерные технологии анализа данных в эконометрике : Монография / Финансовый университет при Правительстве Российской Федерации. - 3-е изд. ; доп. - Москва : Вузовский учебник, 2018. - 587 с. - Дополнительное профессиональное образование. - ISBN 978-5-9558-0275-6. - ISBN 978-5-16-500249-6. - ISBN 978-5-16-006145-0., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=742168&idb=0.

Дополнительная литература:

1. Лемешко Борис Юрьевич. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход : Монография / Новосибирский государственный технический университет. - Москва : ООО "Научно-издательский центр ИНФРА-М", 2015. - 890 с. - Дополнительное профессиональное образование. - ISBN 978-5-16-103267-1., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=594609&idb=0.

Программное обеспечение и Интернет-ресурсы (в соответствии с содержанием дисциплины):

Программное обеспечение

№ п/п Наименование Условия доступа
1. Windows Professional 8.1 Russian Из внутренней сети университета (договор)
2. MS Office 2007 Prof+ Из внутренней сети университета (договор)
3. Среда Anaconda Navigator Программный продукт свободного доступа
4. Jupyter Notebook - командная оболочка для интерактивных вы-числений Программный продукт свободного доступа
5. Google Colab Программный продукт свободного доступа

Интернет-ресурсы

№ п/п	Наименование	Адрес web-страницы
1.	GitHub - веб-сервис для хостинга IT-проектов	https://github.com
2.	Kaggle – сеть специалистов по обработке данных	https://www.kaggle.com/datasets
3.	Habr – ресурс для IT специалистов	https://habr.com/ru/all/

7. Материально-техническое обеспечение дисциплины (модуля)

Учебные аудитории для проведения учебных занятий, предусмотренных образовательной программой, оснащены мультимедийным оборудованием (проектор, экран), техническими средствами обучения, компьютерами.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду.

Программа составлена в соответствии с требованиями ОС ННГУ по направлению подготовки/специальности 39.03.01 - Социология.

Автор(ы): Граница Юлия Валентиновна, кандидат экономических наук, доцент.

Заведующий кафедрой: Болдыревский Павел Борисович, доктор физико-математических наук.

Программа одобрена на заседании методической комиссии от 15.12.2023, протокол № 7.