Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет» (Новосибирский государственный университет, НГУ)

Факультет естественных наук

Согласовано
Декан ФЕН
Резников В.А.
подпись
«10» октября 2020 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

АНАЛИЗ ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ

направление подготовки: 06.04.01 Биология
направленность (профиль): Биология

Форма обучения: очная

Разработчики:

К.б.н. Антонец Д.В.

Руководитель программы:

Д.б.н., проф. Рубцов Н.Б.

Новосибирск, 2020

Содержание

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

Результаты освоения образовательной программы (компетенции)	В результате изучения дисциплины обучающиеся должны:		
	знать	уметь	владеть
ОПК-7 Готовность творчески применять современные компьютерные технологии при сборе, хранении, обработке, анализе и передаче биологической информации для решения профессиональных задач	современные информационно-коммуникационные и интеллектуальные технологии, инструментальные среды, программно-технические платформы для решения профессиональных задач.	обосновывать выбор современных информационно-коммуникационных и интеллектуальных технологий, разрабатывать оригинальные программные средства для решения профессиональных задач.	методами разработки оригинальных программных средств, в том числе с использованием современных информационно-коммуникационных и интеллектуальных технологий, для решения профессиональных задач.

2. Место дисциплины в структуре образовательной программы

Дисциплина *Анализ данных и машинное обучение* является дисциплиной по выбору (Б1.В.ДВ.1.11) и изучается во 2 семестре.

Дисциплина«Анализ данных и машинное обучение»опираетсяна следующие дисциплины данной образовательной программы:

- Основы работы в ОС Linux,
- Программирование на языке Python.

Результатыосвоения дисциплины«Анализ данных и машинное обучение» используются в следующих дисциплинах:

- Учебная практика, ознакомительная практика,
- Производственная практика, научно-исследовательская работа.

3. Трудоемкость дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающегося

Трудоемкость дисциплины – 4з.е. (144 ч)
Форма промежуточной аттестации: Экзамен.

№	Вид деятельности	Количество часов
1	Лекции, ч	36
2	Практические занятия, ч	36

3	Лабораторные занятия, ч	-
4	Занятия в контактной форме, ч из них	76
5	из них аудиторных занятий, ч	72
6	групповая работа с преподавателем, ч	-
7	консультаций, час.	2
8	промежуточная аттестация, ч	2
9	Самостоятельная работа, час.	68
10	Всего, ч	144

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

Лекции (36 часов)

Наименование темы и их содержание	**Часы**
1. Введение в предметную область. Примеры использования методов машинного обучения для решения прикладных задач. Повторение основ программирования на языке Python.	4
2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	4
3. Знакомство с различными методами предобработки данных, описательными статистиками и основными способами визуализации данных, методами снижения размерности. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	4
4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	4
5. Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация. Метод K-средних, DBSCAN и др. Обзор методов кластеризации, реализованных в библиотеке sklearn.	4
6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые. Нормировка данных. Определение переобученности модели. Критерии оценки качества полученных моделей.	4
7. Постановка задачи регрессии. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные наблюдения, анализ остатков. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.	4
8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация. Логистическая регрессия. Решающие деревья. Метрики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой).	4
9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.	4

Практические занятия (36 часов)

Темы практических занятий	**Часы**	**Учебная деятельность**
1. Введение в предметную область. Примеры использования методов машинного обучения для решения прикладных задач. Повторение ос-	4	Примеры использования методов машинного обучения для практических задач. Краткий обзор синтаксиса языка Python. Встроенные операции и функции, типы и структу-

нов программирования на языке Python.		ры данных.
2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	4	Библиотеки NumPy и SciPy. Матрицы. Разреженные матрицы. Индексирование, срезы. Объединение массивов. Библиотека pandas. Запросы к таблицам: выборка строк/столбцов по заданным критериям. Модификация элементов таблицы. Добавление строк/столбцов. Группировка и агрегирование. Объединение таблиц (различные виды join). Многомерные данные: мультииндексы. Операции stack-unstack. Построение сводных таблиц (pivottables).
3. Знакомство с различными методами предобработки данных, описательными статистиками и основными способами визуализации данных, методами снижения размерности. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	4	Описательные статистики. Обзор библиотек matplotlib, seaborn, bokeh. Базовые типы визуализации данных. Знакомство с библиотекой scikit-learn (sklearn). Предобработка данных. Метод главных компонент. Работа с пропущенными значениями.
4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	4	Дальнейшее знакомство студентов с пакетом sklearn. Основные функции. Работа с данными из набора MNIST (рукописные цифры). Работа с синтетическими данными.
5. Обучение на неразмеченных данных. Нормировка данных. Кластеризация. Иерархическая кластеризация. Метод K-средних, DBSCAN и др. Обзор методов кластеризации, реализованных в библиотеке sklearn.	4	Использование методов снижения размерности и методов кластеризации в задаче распознавания рукописных цифр (MNIST). Работа с синтетическими данными.
6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые. Определение переобученности модели. Критерии оценки качества полученных моделей.	4	Примеры задач обучения с учителем. Важность определения целевой метрики качества. Сравнение различных метрик качества моделей. Работа с несбалансированными наборами.
7. Постановка задачи регрессии. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные наблюдения, анализ остатков. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.	4	Объединение алгоритмов, реализованных в sklearn, в цепочки и конвейеры с помощью класса Pipeline. Реализация регрессионных и классификационных моделей с помощью sklearn. Работа с синтетическими данными. Самостоятельная реализация метода градиентного спуска.
8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация. Логистическая регрессия. Решающие деревья. Мет-	4	Реализация классификационных моделей с помощью sklearn. Реализация моделей на основе метода k-ближайших соседей. Метод логистической регрессии. Самостоятельная реализация метода градиентного спуска. Ре-

рики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой).		ализация решающего дерева.
9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.	4	Реализация моделей с помощью метода градиентного бустинга, метода случайного леса. Блендинг и стеккинг. Методы отбора признаков. Оптимизация гиперпараметров.

Самостоятельная работа студентов (68ч)

Перечень занятий на СРС	Объем, час
Самостоятельная работа во время занятий из них:	44
закрепление, обобщение и повторение пройденного учебного материала	10
уточнение и дополнение сведений и знаний, полученных на занятиях	10
выполнение домашнего задания	24
Самостоятельная работа во время промежуточной аттестации из них:	24
подготовка к экзамену	24

5. Перечень учебной литературы

Основная литература

1. Маккинли У. (пер. с англ. СлинкинА.А.), Python и анализ данных // Издательство “ДМК Пресс”, 2015, 482 с.
2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными // Вильямс, 2017, 480 с.

Дополнительная литература (в т.ч. учебная)

1. Бизли Д. Python. Подробный справочник, 4-е издание. Изд-во Символ-Плюс. 2012.
2. ВандерПлас Дж. Python для сложных задач. Наука о данных и машинное обучение. Изд-во “Питер”, 2017, 576 с.
3. Лутц М. Изучаем Python, 5-е издание. Изд-во Символ-Плюс. 2013.
4. Лутц М. Программируем на Python, в 2-х томах, 4-е издание. Изд-во Символ-Плюс. 2011.
5. Рашка С. Python и машинное обучение. Изд-во ДМК Пресс. 2017.
6. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning (second edition) // New York: Springer, 2009.
7. Leskovec J., Rajaraman A., Ullman J.D. Mining of massive datasets // Cambridge University Press, 2014.
8. Bishop C.M. Pattern recognition and machine learning // New York: Springer, 2006.
9. Steele J., Iliinsky N. Beautiful Visualization: Looking at Data through the Eyes of Experts // "O'Reilly Media, Inc.", 2010.

6. Перечень учебно-методических материалов по самостоятельной работе обучающихся

Не используются

7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

№ п/п	Наименование Интернет-ресурса	Краткое описание
1	http://www.machinelearning.ru/	Большая коллекция материалов по машинному обучению на русском языке.
2	http://anaconda.org	Дистрибутив Python с большинством необходимых библиотек.
3	http://scipy.org/	Библиотека для научных вычислений для языка программирования Python.
4	http://pandas.pydata.org/	Библиотека для анализа данных pandas.
5	http://scikit-learn.org/stable/user_guide.html	Документация библиотеки sklearn.
6	http://scikit-learn.org/stable/tutorial/index.html	Примеры решения некоторых задач.
7	http://kaggle.com	Платформа для проведения конкурсов по решению задач машинного обучения. Содержит обучающие ресурсы с примерами решений задач и их обсуждением.
8	http://archive.ics.uci.edu/ml/	Коллекция данных и задач.
9	https://stepik.org/course/67	Курс «Программирование на Python» по основам программирования на языке Python.
10	https://www.coursera.org/learn/machine-learning	Курс по основам машинного обучения от Эндрю ЫНа (AndrewNg). Преподается на английском языке.
11	https://ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie	Курс по основам машинного обучения с использованием Python+ pandas+sklearn. Преподается на русском языке. Преподаватель: Константин Вячеславович Воронцов.

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

8.1 Перечень программного обеспечения

Перечень специализированного программного обеспечения для изучения дисциплины представлен в таблице 8.1.

Специализированное программное обеспечение Таблица 8.1

№	Наименование ПО	Назначение	Место размещения
1	Google Chrome	Интернет браузер	Аудитории 4210, 4211, 4213, 4214, 4218, 4220, 2213, 2221, 3212, 3213, 3218, 3220, 4263 Учебного корпуса №1
2	Anaconda3	Среда разработки приложений	Аудитории 4210, 4211, 4213, 4214, 4218, 4220, 2213, 2221, 3213, 3218, 3220, 4263 Учебного корпуса №1
3		Инструмент для рабо-	Аудитории 4210, 4211, 4213,

	Adobe Acrobat Reader	ты с PDF-файлами	4214, 4218, 4220, 2213, 3213, 4261, 4264, 4263 Учебного корпуса №1
4	Notepad++	Программа для работы с текстовыми файлами	Аудитории 4210, 4211, 4213, 4214, 4218, 4220, 2213, 2221, 3212, 3213 Учебного корпуса №1
5	RStudio	Среда разработки программного обеспечения с открытым исходным кодом для языка программирования R	Аудитории 4261, 4264, 4263, 2221 Учебного корпуса №1

ПО для лиц с ограниченными возможностями здоровья Таблица 8.2

№	Наименование ПО	Назначение	Место размещения
1	JawsforWindows	Программа экранного доступа к системным и офисным приложениям, включая интернет-обозреватели. Информация с экрана считывается вслух, обеспечивая возможность речевого доступа к самому разнообразному контенту. Jaws также позволяет выводить информацию на обновляемый дисплей Брайля. JAWS включает большой набор клавиатурных команд, позволяющих воспроизвести действия, которые обычно выполняются только при помощи мыши.	Ресурсный центр, читальные залы библиотеки НГУ, компьютерные классы (сетевые лицензии)
2	DuxburyBrailleTranslatorv11.3 для Брайлевского принтера	Программа перевода текста в текст Брайля, и печати на Брайлевском принтере	Ресурсный центр
3	"MAGicPro 13" (увеличение+речь)	Программа для людей со слабым зрением и для незрячих людей. Программа позволяет увеличить изображение на экране до 36 крат, есть функция речевого сопровождения	Ресурсный центр, читальные залы библиотеки НГУ

8.2 Информационные справочные системы

1. Полнотекстовые журналы SpringerJournals за 1997-2015 г., электронные книги (2005-2016 гг.), реферативная БД по чистой и прикладной математике zbMATH.
2. Полнотекстовые электронные ресурсы FreedomCollection издательства Elsevier (Нидерланды) (23 предметные коллекции – указать конкретные коллекции)
3. Электронныересурсы Web of Science Core Collection (Thomson Reuters Scientific LLC.), Journal Citation Reports + ESI
4. ЭлектронныеБДJSTOR (США). 15 предметныхколлекций: Arts & Sciences I, II, III, IV, V, VI, VII, VIII, Life Sciences, Health & General Science, Mathematics & Statistics, Ecology & Botany, Language & Literature, Business I, II. – выбратьнужные

5. БД Scopus (Elsevier)

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Для реализации дисциплины *Анализ данных и машинное обучение* используются специальные помещения:

1. Компьютерные классы для проведения групповых и индивидуальных консультаций, текущей и промежуточной аттестации.

Компьютерные классыукомплектованы специализированной мебелью, техническими средствами обучения, служащими для представления учебной информации, и персональными компьютерами.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду НГУ.

Материально-техническое обеспечение образовательного процесса по дисциплине для обучающихся из числа лиц с ограниченными возможностями здоровья осуществляется согласно «Порядку организации и осуществления образовательной деятельности по образовательным программам для инвалидов и лиц с ограниченными возможностями здоровья в Новосибирском государственном университете».

10. Оценочные средства для проведения текущего контроля и промежуточной аттестации по дисциплине

Перечень результатов обучения по дисциплине *Анализ данных и машинное обучение* и индикаторов их достижения представлен в разделе 1.

10.1 Порядок проведения текущего контроля и промежуточной аттестации по дисциплине

Текущая аттестация по дисциплине осуществляется в форме выполнения домашних заданий. Всего предусмотрено 5 домашних заданий. Задания выкладываются на странице курса и в группе курса.Задания нацелены на практическое применение изученных на занятиях методов и алгоритмов. Выполненные задания сдаются в электронном виде. На решение заданий отводится не менее 2 недель. За сдачу задания после 21 дня с даты получения итоговая оценка уменьшается на 10 %. В каждом задании есть теоретическая и практическая часть.

Итоговое задание представляет решение выбранной и согласованной с преподавателем задачи из предметной области курса (анализ данных и машинное обучение).Задание выполняется с использованием языка программированияPython и специализированных библиотек, обязательно должно присутствовать описание использованных данных, указание источника, загрузку данных, разведочный анализ, визуализацию описательных статистик, исследование с помощью методов снижения размерности, описание постановки задачи, построение классификационной или регрессионной модели (в зависимости от типа задачи), обоснование выбора метода и анализ качества полученной модели. Работа должна быть выполнена методологически корректно, без грубых ошибок. Программный код курсовой работы должен исполняться без фатальных ошибок. Студент должен понимать суть выполненной работы и быть готов дать пояснения по программному коду.

Итоговоезадание сдается либо в виде файла *.ipynb – iPythonNotebook, либо в виде презентации с приложением программного кода. Для решения итогового задания допускается использовать язык программирования R.

Промежуточная аттестация по дисциплине проводится в виде экзамена. Оценка выставляется на основе суммы баллов за выполненные домашние задания и за итоговое задания.Суммарное значение баллов, составляющее не менее 85 % от максимального, соответствует оценке «отлично», 70 % – «хорошо», 55 % – «удовлетворительно». Оценки «от-

лично», «хорошо», «удовлетворительно» соответствуют успешному прохождению промежуточной аттестации.

Описание критериев и шкал оценивания индикаторов достижения результатов обучения по дисциплине *Анализ данных и машинное обучение.*

Таблица 10.1

Индикатор	Результат обучения по дисциплине	Оценочное средство
ОПК-7 Готовность творчески применять современные компьютерные технологии при сборе, хранении, обработке, анализе и передаче биоло-гической информации для решения профессиональных задач	*Знает*: современные информационно-коммуникационные и интеллектуальные технологии, инструментальные среды, программно-технические платформы для решения профессиональных задач. *Умеет*: обосновывать выбор современных информационно-коммуникационных и интеллектуальных технологий, разрабатывать оригинальные программные средства для решения профессиональных задач. *Владеет*: методами разработки оригинальных программных средств, в том числе с использованием современных информационно-коммуникационных и интеллектуальных технологий, для решения профессиональных задач.	Практические задания, экзамен

Таблица 10.2

Критерии оценивания результатов обучения	Шкала оценивания
Экзамен: Оценка выставляется на основе суммы баллов за выполненные домашние задания и за итоговое задания.Суммарное значение баллов, составляющее не менее 85 % от максимального	*отлично*
Оценка выставляется на основе суммы баллов за выполненные домашние задания и за итоговое задания.Суммарное значение баллов, составляющее не менее 70 %от максимального	*Хорошо*
Оценка выставляется на основе суммы баллов за выполненные домашние задания и за итоговое задания.Суммарное значение баллов, составляющее не менее 55 % от максимального	*Удовлетворительно*
Оценка выставляется на основе суммы баллов за выполненные домашние задания и за итоговое задания.Суммарное значение баллов, составляющее-менее 55% от максимального	*Неудовлетворительно*