ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85
do@bmstu.ru
do.bmstu.ru

Программа курса
«Junior machine learning engineer: инженер машинного обучения »

Номер	Название темы	Количество часов	Описание темы
1	Обзор библиотеки sklearn	1,5	Библиотека scikit-learn (sklearn), назначение, разделы, способы работы, импорт библиотеки в python. Практическая работа (0,5 часа). Работа с библиотекой scikit-learn (sklearn). Самостоятельная работа (0,5 часа)
2	Метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки	4	Методы уменьшения размерности PCA и t-SNE, даны определения линейно-разделимой и неразделимой выборки, в каких датасетах и в каких данных необходимо уменьшение размерность. Практическая работа (1 час). Изменение размерности датасета. Самостоятельная работа (2 часа).
3	Кластеризация. Метод k-means, c-means	4	Алгоритмы кластеризации k-mean и c-means, как примеры обучения без учителя, основные особенности. Написание кода алгоритма на python. Практическая работа (1 час). Кластеризация датасета используя алгоритмы k-means, c-means. Самостоятельная работа (2 часа).
4	Иерархическая кластеризация - hierarchical clustering. Алгоритм кластеризации DBSCAN	3	Алгоритмы иерархической кластеризации и алгоритм DBSCAN. Преимущества и недостатки. Практическая работа (1 час). Кластеризация научных патентов с применением hierarchical clustering и DBSCAN. Самостоятельная работа (1 час).
5	Ключевые задачи в подготовке датасетов и их важность	1	Подготовка датасетов, проверка на полноту, оценка пропущенных значений, валидация данных и источников, достоверность, многообразие. Самостоятельная работа (0,5 часа).
6	Разбалансированные датасеты и методы балансировки	2	Разбалансированный датасет и балансировка, миноритарный класс, мажоритарный класс. Применение методов увеличения миноритарного класса (upsampling) и уменьшения мажоритарного класса (downsampling). Практическая работа (0,5 часа). Применение методов балансировки датасетов. Самостоятельная работа (1 час).
7	Библиотека Beautifulsoup. Парсинг данных из html страниц	3,5	Метод и реализация парсинга (сбора) данных из открытых источников в интернете с применением библиотеки beautifulsoup. Будет предложен вариант навигации по коду html страниц. Практическая работа (1 час). Выполнить парсинг двух страниц с сайта https://zakupki.gov.ru/ по каждой закупке. Самостоятельная работа (2 часа).
8	Обработка категориальных признаков. LabelEncoder, One Hot encoding	2	Категориальные и числовые признаки, а также методы обработки категориальных признаков LaberlEncoder, One Hot encoding. Будут разобраны условия. когда оптимально применять методы обработки категориальных признаков. Практическая работа (0,5 часа). Загрузить и собрать датасет (датасет описывает классические автомобили), определить категориальные признаки, применить методы LabelEncoder, One Hot Encoding. Самостоятельная работа (1 часа).
	Полная и условная		Понятия полной и условной вероятности, а также теорема Байеса,

9	вероятность, теорема Байеса	1	зависимые и независимые события. Самостоятельная работа (0,5 часа)
10	Байесовский вероятностный классификатор	3	Вероятностные классификаторы Байеса (Gaussian Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes, Bernoulli Naive Bayes, Categorical Naive Bayes) для решения задач классификации. Практическая работа (1 час). Обучить два Байесовских классификатора, спрогнозировать вероятность возникновения лесных пожаров. Выполнить прогноз на проверочных данных. Самостоятельная работа (1,5 часа).
11	Метрики классификации. Матрица ошибок (Confusion -matrix) Precision, recall, f1. ROC-AUC	2	Метрики ошибок при решении задач классификации, даны определения для метрик precision, recall, f1-мера, построение ROC Кривой. Практическая работа (0,5 часа). Комплексная оценка работы алгоритма по набору метрик. Самостоятельная работа (0,5 часа).
12	Кросс-валидация. Особенности применения	1	Понятие кросс-валидации, преимущества при оценке и проверке качества алгоритмов машинного обучения. Самостоятельная работа (0,5 часа).
13	Метод ближайших соседей k-NN. Метрики подсчета расстояния. Плюсы и минусы алгоритма	3	Алгоритм машинного обучения - метод ближайших соседей (k-NN), для решения задач классификации. Область решаемых задач. Плюсы и минусы алгоритма. Практическая работа (1 час). Обучить алгоритм k-NN, используя две разные метрики близости. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Написать выводы. Самостоятельная работа (1 час).
14	Метод опорных векторов (SVM). Линейно разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма	3	Алгоритм - метод опорных векторов, проблема линейно не разделимой выборки и методы её решения. Область решаемых задач, плюсы и минусы алгоритма. Практическая работа (1 час). Обучить алгоритм SVM. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Самостоятельная работа (1 час).
15	Линейная регрессия. Логистическая регрессия (4 часа)	4	Основные термины и понятия линейной регрессии, логистической регрессии, регуляризации, смещения и дисперсии (разброса). Практическая работа (1 час). Обучить алгоритм линейной регрессии для прогнозирования значений. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Самостоятельная работа (2 часа).
16	Метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка	2	Метрики оптимизации и ошибок для задач регрессии, такие как метод наименьших квадратов, средняя абсолютная и квадратичная ошибки, будут даны определения. Практическая работа (0,5 часов). Решение задач методом наименьших квадратов. Самостоятельная работа (0,5 часов).
17	Решающие деревья (Decision tree)	4	Алгоритм решающих деревьев (Decision tree), для решения прикладных задач, области решаемых задач, плюсы и минусы алгоритма. Практическая работа (1 час). Обучить алгоритм Decision tree. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Самостоятельная работа (2 часа).
18	Случайный лес (Random forest)	3	Алгоритм случайного леса (Random forest), ключевые отличия от decision tree, области решаемых задач, плюсы и минусы алгоритма. Практическая работа (1 час). Обучить алгоритм Random forest. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Самостоятельная работа (1,5 часа).
			Ансамблевые алгоритмы для повышения точности. bagging - параллельный, boosting - последовательный, stacking - совместный

19	Ансамбли алгоритмов. Bagging, boosting, stacking	3	запуск алгоритмов. Области решаемых задач, плюсы и минусы подхода. Практическая работа (0,5 часа). Обучить алгоритм. Применить Boosting и bagging ансамбль. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Написать выводы. Самостоятельная работа (2 часа).
20	Итоговая аттестация	2	Тестирование

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Первый проректор –
проректор по учебной работе
МГТУ им. Н.Э. Баумана
_____ Б.В. Падалкин
«13» 12 _____ 2021 г.

Дополнительное профессиональное образование

ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА ПОВЫШЕНИЯ КВАЛИФИКАЦИИ
«Junior machine learning engineer: инженер машинного обучения»

Регистрац. № 05.22.21.12.2

Москва, 2021

АВТОР ПРОГРАММЫ:

Доцент
Центра НТИ «Цифровое материаловедение:
новые материалы и вещества» _____ В.С. Тынченко

СОГЛАСОВАНО:

Руководитель стратегического проекта «Bauman U2U» _____ А.С. Комшин

Начальник УСП
МГТУ им. Н.Э. Баумана _____ Т.А. Гузева

Директор ИСОТ
МГТУ им. Н.Э. Баумана _____ В.Г. Брекалов

Заместитель директора
Центра НТИ «Цифровое материаловедение:
новые материалы и вещества» _____ М.В. Стоянова

Оглавление

1. Общая характеристика дополнительной профессиональной программы (ДПП)

Программа подготовлена на основе:

- Федерального закона от 29 декабря 2012 года № 273-ФЗ «Об образовании в Российской Федерации»;
- приказа Минобрнауки России от 23 августа 2017 года № 816 «Об утверждении Порядка применения организациями, осуществляющими образовательную деятельность, электронного обучения, дистанционных образовательных технологий при реализации образовательных программ»;
- требований Приказа Минобрнауки России от 01.07.2013 года № 499 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным профессиональным программам»;
- методических рекомендаций-разъяснений Минобрнауки России по разработке дополнительных профессиональных программ на основе профессиональных стандартов от 22 апреля 2015 года № ВК-1030/06.

Реализация программы дополнительного профессионального образования направлена на получение новой(-ых) компетенции(-ий), необходимой(-ых) для профессиональной деятельности, в рамках реализации программы стратегического академического лидерства «Приоритет-2030».

1.1. Цель ДПП

Сформировать у обучающихся компетенции в области создания информационных технологий нового поколения, обеспечивающих экономически эффективное извлечение полезной информации из больших объемов разнообразных данных путем высокой скорости их сбора, обработки и анализа, и применение этих технологий в информационно-аналитической деятельности, в системах управления и принятия решений, а также для разработки на их основе новых продуктов и услуг.

1.2. Планируемые результаты обучения

Планируемые результаты обучения по ДПП:

- освоение профессиональных компетенций в процессе изучения перечисленных тем в учебном плане;
- успешное освоение программы повышения квалификации;
- успешное прохождение итоговой аттестации (зачет).

Обучающимся, успешно прошедшим обучение, выполнившим текущие контрольные задания и выдержавшим предусмотренное учебным планом итоговую аттестацию, выдается удостоверение о повышении квалификации по ДПП «Junior machine learning engineer: инженер машинного обучения».

1.3. Дополнительные характеристики ДПП

Перечень профессиональных компетенций в рамках имеющейся квалификации, качественное изменение которых осуществляется в результате обучения, определены в Приказом Министерства труда и социальной защиты Российской Федерации от 06.07.2020 №405н об утверждении профессионального стандарта «Специалист по большим данным».

Вид профессиональной деятельности:

- Создание и применение технологий больших данных (Код 06.042).

Обобщенные трудовые функции:

- Управление этапами жизненного цикла методологической и технологической инфраструктуры анализа больших данных в организации (ОТФ 06.042_B).

Трудовые функции:

- Управление разработкой продуктов, услуг и решений на основе больших данных (ТФ 06.042_B/05.7).

1.4. Перечень профессиональных компетенций в рамках имеющейся квалификации, качественное изменение которых осуществляется в результате обучения

Данная программа направлена на совершенствование и (или) получение новой(-ых) компетенции, необходимой для профессиональной деятельности, и (или) повышение профессионального уровня в рамках имеющейся квалификации.

При определении профессиональных компетенций на основе профессиональных стандартов, МГТУ им. Н.Э.Баумана осуществляет выбор профессиональных стандартов, соответствующих профессиональной деятельности выпускников, из числа указанных в приложении к ФГОС ВО и (или) иных профессиональных стандартов, соответствующих профессиональной деятельности выпускников.

Получаемые компетенции базируются на основании Приказа Минобрнауки России от 19 сентября 2017 г. № 932 «Об утверждении Федерального государственного образовательного стандарта высшего образования – магистратура по направлению подготовки 09.04.04 Программная инженерия».

Перечень компетенций:

ПК-1. Способен создавать информационные системы, понимание существующих подходов к верификации моделей программного обеспечения.

ОПК-5. Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем.

1.5. Соответствие видов деятельности профессиональным компетенциям и их составляющих

Профессиональные компетенции	Практический опыт	Умения	Знания
Управление разработкой продуктов, услуг и решений на основе больших данных (B/05.7)			
ПК-1. Способен создавать информационные системы, понимание существующих подходов к верификации моделей программного обеспечения	Разработка моделей данных, адаптированных к технологиям больших данных	Пользоваться методами и инструментами получения, хранения, передачи, обработки больших данных	Этапы жизненного цикла больших данных
ОПК -5. Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем			

2. Учебный план ДПП

2.1. Категория слушателей ДПП

Имеющаяся квалификация (требования к обучающимся) – к освоению ДПП допускаются лица, соответствующего профессиональному стандарту уровню образования или получающие соответствующий уровень образования (специалитет, магистратура).

2.2. Общая трудоёмкость программы, аудиторная и самостоятельная работа

Общая трудоемкость программы 52 академических часа, из них 27 часов аудиторной работы, 23 часа самостоятельной работы и 2 часа итоговой аттестации.

2.3. Форма обучения

Форма обучения по ДПП – очная, с применением электронного обучения и дистанционных образовательных технологий.

2.4. Учебный план

ДПП «Junior machine learning engineer: инженер машинного обучения» реализуется одним модулем.

№ п/п	Наименование темы, модуля	Форма контроля	Всего, акад. час*	В том числе		
				Лекции	Практические занятия	Самостоятельная работа
1.	Обзор библиотеки sklearn	тест	1,5	0,5	0,5	0,5
2.	Метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки.	устный опрос	4	1	1	2
3.	Кластеризация. Метод k-means, c-means	устный опрос	4	1	1	2
4.	Иерархическая кластеризация - hierarchical clustering. Алгоритм кластеризации DBSCAN	устный опрос	3	1	1	1
5.	Ключевые задачи в подготовке датасетов и их важность	тест	1	0,5	-	0,5
6.	Разбалансированные датасеты и методы балансировки	тест	2	0,5	0,5	1
7.	Библиотека Beautifulsoup. Парсинг данных из html страниц	тест	3,5	0,5	1	2
8.	Обработка категориальных признаков. LabelEncoder, One Hot encoding	тест	2	0,5	0,5	1
9.	Полная и условная вероятность, теорема Байеса	устный опрос	1	0,5	-	0,5
10.	Байесовский вероятностный классификатор	устный опрос	3	0,5	1	1,5
11.	Метрики классификации.	устный опрос	2	1	0,5	0,5

	Матрица ошибок (Confusion - matrix) Precision, recall, f1. ROC-AUC					
12.	Кросс-валидация. Особенности применения	устный опрос	1	0,5	-	0,5
13.	Метод ближайших соседей k-NN. Метрики подсчета расстояния. Плюсы и минусы алгоритма	устный опрос	3	1	1	1
14.	Метод опорных векторов (SVM). Линейно разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма	устный опрос	3	1	1	1
15.	Линейная регрессия. Логистическая регрессия	устный опрос	4	1	1	2
16.	Метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка	устный опрос	2	1	0,5	0,5
17.	Решающие деревья (Decision tree)	устный опрос	4	1	1	2
18.	Случайный лес (Random forest)	устный опрос	3	0,5	1	1,5
19.	Ансамбли алгоритмов. Bagging, boosting, stacking	устный опрос	3	0,5	0,5	2
20.	Итоговая аттестация	зачет	2	-	-	-
	ИТОГО	-	52	14	13	23

*академический час составляет 45 минут

3. Календарный учебный график

№ п/п	Наименование темы, модуля	1 день	2 день	3 день	4 день	5 день	6 день
1.	Обзор библиотеки sklearn						
2.	Метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки.						
3.	Кластеризация. Метод k-means, c-means						
4.	Иерархическая кластеризация - hierarchical clustering. Алгоритм кластеризации DBSCAN						
5.	Ключевые задачи в подготовке датасетов и их важность						
6.	Разбалансированные датасеты и методы балансировки						
7.	Библиотека Beautifulsoup. Парсинг данных из html страниц						
8.	Обработка категориальных признаков. LabelEncoder, One Hot encoding						
9.	Полная и условная вероятность, теорема Байеса						
10.	Байесовский вероятностный классификатор						
11.	Метрики классификации. Матрица ошибок (Confusion -matrix) Precision, recall, f1. ROC-AUC						
12.	Кросс-валидация. Особенности применения						
13.	Метод ближайших соседей k-NN. Метрики подсчета расстояния. Плюсы и минусы алгоритма						
14.	Метод опорных векторов (SVM). Линейно						

	разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма						
15.	Линейная регрессия. Логистическая регрессия						
16.	Метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка						
17.	Решающие деревья (Decision tree)						
18.	Случайный лес (Random forest)						
19.	Ансамбли алгоритмов. Bagging, boosting, stacking						
20.	Итоговая аттестация						

№ п/п	Наименование темы, модуля	7 день	8 день	9 день	10 день	11 день
1.	Обзор библиотеки sklearn					
2.	Метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки.					
3.	Кластеризация. Метод k-means, c-means					
4.	Иерархическая кластеризация - hierarchical clustering. Алгоритм кластеризации DBSCAN					
5.	Ключевые задачи в подготовке датасетов и их важность					
6.	Разбалансированные датасеты и методы балансировки					
7.	Библиотека Beautifulsoup. Парсинг данных из html страниц					
8.	Обработка категориальных признаков. LabelEncoder, One Hot encoding					
9.	Полная и условная вероятность, теорема Байеса					
10.	Байесовский вероятностный классификатор					
11.	Метрики классификации. Матрица ошибок (Confusion					

	-matrix) Precision, recall, f1. ROC-AUC					
12.	Кросс-валидация. Особенности применения					
13.	Метод ближайших соседей k-NN. Метрики подсчета расстояния. Плюсы и минусы алгоритма					
14.	Метод опорных векторов (SVM). Линейно разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма					
15.	Линейная регрессия. Логистическая регрессия					
16.	Метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка					
17.	Решающие деревья (Decision tree)					
18.	Случайный лес (Random forest)					
19.	Ансамбли алгоритмов. Bagging, boosting, stacking					
20.	Итоговая аттестация					Зачет

Рекомендованный срок освоения ДПП — 11 дней.
Фактическое расписание занятий утверждается при заключении договора с обучающимися или при формировании группы.

4. Рабочая программа ДПП

4.1. Рабочая программа модуля
4.1.1. Цель изучения модуля: сформировать у обучающихся компетенции в области создания информационных технологий нового поколения, обеспечивающих экономически эффективное извлечение полезной информации из больших объемов разнообразных данных путем высокой скорости их сбора, обработки и анализа, и применение этих технологий в информационно-аналитической деятельности, в системах управления и принятия решений, а также для разработки на их основе новых продуктов и услуг.
4.1.2. Задача изучения модуля: изучить средства разработки систем, использующих в своей работе принципы машинного обучения.
4.1.3. Планируемые результаты обучения
Процесс изучения раздела направлен на формирование следующих компетенций

Код/ наименование компетенции	Перечень планируемых результатов обучения по модулю	Методы и формы обучения, способствующие формированию и развитию компетенции
ПК-1	**Знать:**	Методы обучения:

	Этапы жизненного цикла больших данных. **Уметь:** Пользоваться методами и инструментами получения, хранения, передачи, обработки больших данных. **Владеть:** Разработка моделей данных, адаптированных к технологиям больших данных.	Активные, пассивные, интерактивные. Формы обучения: лекция; практическое занятие; самостоятельная работа.
ОПК-5	**Знать:** Этапы жизненного цикла больших данных. **Уметь:** Пользоваться методами и инструментами получения, хранения, передачи, обработки больших данных. **Владеть:** Разработка моделей данных, адаптированных к технологиям больших данных.	Методы обучения: Активные, пассивные, интерактивные. Формы обучения: лекция; практическое занятие; самостоятельная работа.

4.1.4 Содержание курса

Тема 1. Обзор библиотеки sklearn (1,5 часа)

Лекции (0,5 часа). В рамках данной темы будет рассмотрена библиотека scikit-learn (sklearn), назначение, разделы, способы работы, импорт библиотеки в python.

Практическая работа (0,5 часа). Работа с библиотекой scikit-learn (sklearn).

Самостоятельная работа (0,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Обзор библиотеки sklearn	Библиотека sklearn	Проработка дополнительных источников информации	Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных / Силен Д., Мейсман А., Али М.; пер. с англ. Матвеев Е. – СПб.: Питер, 2020. – 334 с. (https://library.bmstu.ru/Catalog/Details/544371)	тест

Тема 2. Метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки (4 часа)

Лекции (1 час). В рамках данной темы будут рассмотрены методы уменьшения размерности PCA и t-SNE, даны определения линейно-разделимой и неразделимой выборки, в каких датасетах и в каких данных необходимо уменьшение размерность.

Практическая работа (1 час). Изменение размерности датасета.

Самостоятельная работа (2 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки	Размерность датасета	Проработка дополнительных источников информации	Шелухин О.И., Ерохин С.Д., Полковников М.В. Технологии машинного обучения в сетевой безопасности/ Шелухин О.И., Ерохин С.Д., Полковников М.В.; ред. Шелухин О.И. – М.: Горячая линия-Телеком, 2021. – 359 с. (https://library.bmstu.ru/Catalog/Details/555230)	устный опрос

Тема 3. Кластеризация. Метод k-means, c-means (4 часа)

Лекции (1 час). В данной теме будут рассмотрены два алгоритма кластеризации k-mean и c-means, как примеры обучения без учителя, основные особенности. Написание кода алгоритма на python.

Практическая работа (1 час). Кластеризация датасета используя алгоритмы k-means, c-means.

Самостоятельная работа (2 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Кластеризация. Метод k-means, c-means	Алгоритмы k-means, c-means	Проработка дополнительных источников информации	Сузи Р. Python. В подлиннике: Наиболее полное руководство / Сузи Р. – СПб.: БХВ-Петербург, 2002. – 747 с. (https://library.bmstu.ru/Catalog/Details/84709)	устный опрос

Тема 4. Иерархическая кластеризация - hierarchical clustering. Алгоритм кластеризации DBSCAN (3 часа)

Лекции (1 час). В рамках данной темы будут рассмотрены алгоритмы иерархической кластеризации и алгоритм DBSCAN. Преимущества и недостатки.

Практическая работа (1 час). Кластеризация научных патентов с применением hierarchical clustering и DBSCAN.

Самостоятельная работа (1 час).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Иерархическая кластеризация - hierarchical clustering. Алгоритм кластеризации DBSCAN	Основы кластеризации	Проработка дополнительных источников информации	Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение / Плас Дж. Вандер; пер. с англ. Пальти И. – СПб.: Питер, 2020. – 572 с. (https://library.bmstu.ru/Catalog/Details/550326)	устный опрос

Тема 5. Ключевые задачи в подготовке датасетов и их важность (1 час)

Лекции (0,5 часа). В рамках данной темы будут рассмотрены важные моменты в подготовке датасетов, такие как проверка на полноту, оценка пропущенных значений, валидация данных и источников, достоверность, многообразие.

Самостоятельная работа (0,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Ключевые задачи в подготовке датасетов и их важность	Подготовка данных	Проработка дополнительных источников информации	Круз Р.Л. Структуры данных и проектирование программ: [учеб. пособие] / Круз Р.Л.; пер. 3-го англ. изд. Финогенов К.Г. – М.: БИНОМ. Лаборатория знаний, 2017. – 765 с. (https://library.bmstu.ru/Catalog/Details/476469)	тест

Тема 6. Разбалансированные датасеты и методы балансировки (2 часа)

Практическая работа (0,5 часа). В рамках данной темы будут рассмотрены понятия разбалансированного датасета и балансировка, миноритарный класс, мажоритарный класс. Применение методов увеличения миноритарного класса (upsampling) и уменьшения мажоритарного класса (downsampling).

Практическая работа (0,5 часа). Применение методов балансировки датасетов.

Самостоятельная работа (1 час).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Разбалансированные датасеты и методы балансировки	Балансировка данных	Проработка дополнительных источников информации	Круз Р.Л. Структуры данных и проектирование программ: [учеб. пособие] / Круз Р.Л.; пер. 3-го англ. изд. Финогенов К.Г. – М.: БИНОМ. Лаборатория знаний, 2017. – 765 с. (https://library.bmstu.ru/Catalog/Details/476469)	тест

Тема 7. Библиотека Beautifulsoup. Парсинг данных из html страниц (3,5 часа)

Практическая работа (0,5 часа). В рамках данной темы будет рассмотрен метод и реализация парсинга (сбора) данных из открытых источников в интернете с применением библиотеки beautifulsoup. Будет предложен вариант навигации по коду html страниц.

Практическая работа (1 час). Выполнить парсинг двух страниц с сайта https://zakupki.gov.ru/ по каждой закупке.

Самостоятельная работа (2 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Библиотека Beautifulsoup. Парсинг данных из html страниц	Парсинг данных	Проработка дополнительных источников информации	Круз Р.Л. Структуры данных и проектирование программ: [учеб. пособие] / Круз Р.Л.; пер. 3-го англ. изд. Финогенов К.Г. – М.: БИНОМ. Лаборатория знаний, 2017. – 765 с. (https://library.bmstu.ru/Catalog/Details/476469)	тест

Тема 8. Обработка категориальных признаков. LabelEncoder, One Hot encoding (2 часа)

Практическая работа (0,5 часа). В данной теме будут рассмотрены понятия категориальные и числовые признаки, а также методы обработки категориальных признаков LaberlEncoder, One Hot encoding. Будут разобраны условия. когда оптимально применять методы обработки категориальных признаков.

Практическая работа (0,5 часа). Загрузить и собрать датасет (датасет описывает классические автомобили), определить категориальные признаки, применить методы LabelEncoder, One Hot Encoding.

Самостоятельная работа (1 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Обработка категориальных признаков. LabelEncoder, One Hot encoding	Категориальные признаки данных	Проработка дополнительных источников информации	Круз Р.Л. Структуры данных и проектирование программ: [учеб. пособие] / Круз Р.Л.; пер. 3-го англ. изд. Финогенов К.Г. – М.: БИНОМ. Лаборатория знаний, 2017. – 765 с. (https://library.bmstu.ru/Catalog/Details/476469)	тест

Тема 9. Полная и условная вероятность, теорема Байеса (1 час)

Лекции (0,5 часа). В рамках данной темы будут рассмотрены понятия полной и условной вероятности, а также теорема Байеса, зависимые и независимые события.

Самостоятельная работа (0,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Полная и условная вероятность, теорема Байеса	Теорема Байеса	Проработка дополнительных источников информации	Галкин С.В., Панов В.Ф., Петрухина О.С. Краткий курс теории вероятностей: учеб. пособие / Галкин С.В., Панов В.Ф., Петрухина О.С.; МГТУ им. Н.Э. Баумана. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2007. – 54 с. (https://library.bmstu.ru/Catalog/Details/178041)	устный опрос

Тема 10. Байесовский вероятностный классификатор (3 часа)

Лекции (0,5 часа). В данной теме будут рассмотрены вероятностные классификаторы Байеса (Gaussian Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes, Bernoulli Naive Bayes, Categorical Naive Bayes) для решения задач классификации.

Практическая работа (1 час). Обучить два Байесовских классификатора, спрогнозировать вероятность возникновения лесных пожаров. Выполнить прогноз на проверочных данных.

Самостоятельная работа (1,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Байесовский вероятностный классификатор	Задачи классификации	Проработка дополнительных источников информации	Хливненко Л.В., Пятакович Ф.А. Практика нейросетевого моделирования: учебное пособие / Хливненко Л.В., Пятакович Ф.А. – 2-е изд., стер. – СПб.: Лань, 2021. – 196 с. (https://library.bmstu.ru/Catalog/Details/556123)	устный опрос

Тема 11. Метрики классификации. Матрица ошибок (Confusion -matrix) Precision, recall, f1. ROC-AUC (2 часа)

Лекции (1 час). В рамках данной темы будут рассмотрены метрики ошибок при решении задач классификации, даны определения для метрик precision, recall, f1-мера, построение ROC Кривой.

Практическая работа (0,5 часа). Комплексная оценка работы алгоритма по набору метрик.

Самостоятельная работа (0,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Метрики классификации. Матрица ошибок (Confusion -matrix) Precision, recall, f1. ROC-AUC	Метрики классификации	Проработка дополнительных источников информации	Грановская Р.М., Березная И.Я. Интуиция и искусственный интеллект / Грановская Р.М., Березная И.Я.; Ленинградский гос. ун-т. – Л.: Изд-во Ленинградского ун-та, 1991. – 268 с. (https://library.bmstu.ru/Catalog/Details/503904)	устный опрос

Тема 12. Кросс-валидация. Особенности применения (1 час)

Лекции (0,5 часа). В рамках данной темы будет рассмотрено понятие кросс-валидации, преимущества при оценке и проверке качества алгоритмов машинного обучения.

Самостоятельная работа (0,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Кросс-валидация. Особенности применения	Оценка качества алгоритмов машинного обучения	Проработка дополнительных источников информации	Шумский С.А. Машинный интеллект. Очерки по теории машинного обучения и искусственного интеллекта / Шумский С.А.; Московский физико-технический ин-т (национальный исследовательский ун-т). – М.: РИОР: Инфра-М, 2021. – 339 с. (https://library.bmstu.ru/Catalog/Details/554838)	устный опрос

Тема 13. Метод ближайших соседей k-NN. Метрики подсчета расстояния. Плюсы и минусы алгоритма (3 часа)

Лекции (1 час). В рамках данной темы будет рассмотрен алгоритм машинного обучения - метод ближайших соседей (k-NN), для решения задач классификации. Область решаемых задач. Плюсы и минусы алгоритма.

Практическая работа (1 час). Обучить алгоритм k-NN, используя две разные метрики близости. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Написать выводы.

Самостоятельная работа (1 час).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Метод ближайших соседей k-NN. Метрики подсчета расстояния. Плюсы и минусы алгоритма	Метод ближайших соседей	Проработка дополнительных источников информации	Берикашвили В.Ш., Оськин С.П. Статистическая обработка данных, планирование эксперимента и случайные процессы: учебное пособие для вузов / Берикашвили В.Ш., Оськин С.П. - 2-е изд.,	устный опрос

			испр. и доп. – М.: Юрайт, 2021. – 163 с. (https://library.bmstu.ru/Catalog/Details/555713)	

Тема 14. Метод опорных векторов (SVM). Линейно разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма (3 часа)

Лекции (1 час). В данной теме будет рассмотрен алгоритм - метод опорных векторов, проблема линейно не разделимой выборки и методы её решения. Область решаемых задач, плюсы и минусы алгоритма.

Практическая работа (1 час). Обучить алгоритм SVM. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели.

Самостоятельная работа (1 час).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Метод опорных векторов (SVM). Линейно разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма	Метод опорных векторов	Проработка дополнительных источников информации	Вьюгин В.В. Элементы математической теории машинного обучения: учеб. пособие для вузов / Вьюгин В.В.; Моск. физико-техн. ин-т (гос. ун-т), РАН. Ин-т проблем передачи информации им. А.А. Харкевича. – М.: МФТИ - ИППИ РАН, 2010. – 231 с. (https://library.bmstu.ru/Catalog/Details/229808)	устный опрос

Тема 15. Линейная регрессия. Логистическая регрессия (4 часа)

Лекции (1 час). В рамках данной темы будут рассмотрены основные термины, и понятия линейной регрессии, логистической регрессии, регуляризации, смещения и дисперсии (разброса).

Практическая работа (1 час). Обучить алгоритм линейной регрессии для прогнозирования значений. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели.

Самостоятельная работа (2 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Линейная регрессия. Логистическая регрессия	Виды регрессии	Проработка дополнительных источников информации	Меженная Н.М. Основы теории вероятностей и математической статистики: курс лекций / Меженная Н.М.; МГТУ им. Н.Э. Баумана. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2016. – 108 с. (https://library.bmstu.ru/Catalog/Details/465260)	устный опрос

Тема 16. Метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка (2 часа)

Лекции (1 час). В рамках данной темы будут рассмотрены метрики оптимизации и ошибок для задач регрессии, такие как метод наименьших квадратов, средняя абсолютная и квадратичная ошибки, будут даны определения.

Практическая работа (0,5 часов). Решение задач методом наименьших квадратов.

Самостоятельная работа (0,5 часов).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка	Метод наименьших квадратов	Проработка дополнительных источников информации	Амосов А.А., Дубинский Ю.А., Копченова Н.В. Вычислительные методы: учеб. пособие / Амосов А.А., Дубинский Ю.А., Копченова Н.В. - 4-е изд., стер. – СПб.: Лань, 2014. – 671 с. (https://library.bmstu.ru/Catalog/Details/379247)	устный опрос

Тема 17. Решающие деревья (Decision tree) (4 часа)

Лекции (1 час). В рамках данной темы будут рассмотрен алгоритм решающих деревьев (Decision tree), для решения прикладных задач, области решаемых задач, плюсы и минусы алгоритма.

Практическая работа (1 час). Обучить алгоритм Decision tree. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели.

Самостоятельная работа (2 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Решающие деревья (Decision tree)	Решающие деревья	Проработка дополнительных источников информации	Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознования образов и анализа разнотипной информации / Лбов Г.С., Бериков В.Б. – Новосибирск: Изд-во Ин-та математики, 2005. – 217 с. (https://library.bmstu.ru/Catalog/Details/126725)	устный опрос

Тема 18. Случайный лес (Random forest) (3 часа)

Лекции (0,5 часа). В данной теме будет рассмотрен алгоритм случайного леса (Random forest), ключевые отличия от decision tree, области решаемых задач, плюсы и минусы алгоритма.

Практическая работа (1 час). Обучить алгоритм Random forest. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели.

Самостоятельная работа (1,5 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Случайный лес (Random forest)	Случайный лес	Проработка дополнительных источников информации	Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознования образов и анализа разнотипной информации / Лбов Г.С., Бериков В.Б. – Новосибирск: Изд-во Ин-та математики, 2005. – 217 с. (https://library.bmstu.ru/Catalog/Details/126725)	устный опрос

Тема 19. Ансамбли алгоритмов. Bagging, boosting, stacking (3 часа)

Лекции (0,5 часа). В рамках данной темы будут рассмотрены ансамблевые алгоритмы для повышения точности. bagging - параллельный, boosting -

последовательный, stacking - совместный запуск алгоритмов. Области решаемых задач, плюсы и минусы подхода.

Практическая работа (0,5 часа). Обучить алгоритм. Применить Boosting и bagging ансамбль. Выполнить прогноз на проверочных данных. Выполнить прогноз на проверочных данных. Снять метрики и ошибки модели. Написать выводы.

Самостоятельная работа (2 часа).

Наименование темы	Дидактические единицы, вынесенные на самостоятельное изучение	Формы самостоятельной работы	Учебно-методическое обеспечение	Форма контроля
Ансамбли алгоритмов. Bagging, boosting, stacking	Ансамбли алгоритмов	Проработка дополнительных источников информации	Рассел С., Норвиг П. Искусственный интеллект. Современный подход / Рассел С., Норвиг П.; пер. с англ. и ред. Птицын К.А. - 2-е изд. – М.: Изд. дом «Вильямс», 2018. – 1407 с. (https://library.bmstu.ru/Catalog/Details/518925)	устный опрос

4.1.5. Оценочное средство для текущего контроля (темы для подготовки к тесту и устному опросу):

Тема 1. Тест: библиотека sklearn, назначение, разделы, способы работы (тест считается сданным при верном ответе на 3 из 5 вопросов).

1. Что такое PyTorch?

A. Современная библиотека глубокого обучения от Facebook
B. Современная библиотека глубокого обучения от Google
C. Библиотека поддержки работы Tensorflow разработанная Google
D. Библиотека, созданная на основе Tensorflow для глубокого обучения

2. Выберете основные характеристики PyTorch (Несколько вариантов ответа)

A. Простой интерфейс
B. Использование Python
C. Вычислительные графы
D. Сложный интерфейс
E. Отсутствие графов
F. Использование быстродействующих функций вычисления

3. Что из списка является преимуществами PyTorch? (Несколько вариантов ответа)

A. Код легко отлаживать и понимать
B. Большой функционал настроек
C. Много метрик

D. Интеграция с numpy
E. Сложные структуры сетей

4. Какой фреймворк является старше?

A. PyTorch является более старшим, чем Tensorflow
B. Tensorflow является более старшим, чем PyTroch
C. PyTorch и Tensorflow являются ровесниками

5. Можно ли строить граф в PyTroch параллельно с его обучением?

A. Да, это основная особенность PyTroch
B. Да, но не во всех случаях
C. Нет, такой функции в PyTorch нет

Тема 2. Устный опрос: метод главных компонент PCA. Метод t-SNE для линейно разделимой выборки.
Тема 3. Устный опрос: кластеризация. к-средних (k-means), c-средних (c-means).
Тема 4. Устный опрос: иерархическая кластеризация - hierarchical clustering. Алгоритм пространственной кластеризации DBSCAN.
Тема 5. Тест: ключевые задачи в подготовке датасетов (тест считается сданным при верном ответе на 4 из 5 вопросов).

1. Что такое Data Quality?

A. Характеристика, показывающая насколько данные подходят для чтения человеком
B. Характеристика, показывающая степень пригодности данных к использованию.
C. Характеристика, показывающая насколько данные подходят для чтения машиной
D. Характеристика, показывающая в скольких задачах сразу можно использовать эти данные

2. Какой критерий говорит о высоком качестве данных?

A. Пригодность для использования в операциях, принятии решений и планировании
B. Разнообразие данных (Большое количество фичей)
C. Большое количество данных с большим их разнообразием
D. Хорошая читаемость и человеком, и машиной

3. Если данные правильно представляют реальную конструкцию, к которой они относятся, то скорее всего они являются…

A. Высококачественными
B. Данными среднего качества
C. Низкокачественными

4. Почему релевантность имеет значение как характеристика качества данных?

A. Это позволяет понять если данные наполнены дубликатами

B. Если собирать слишком узконаправленную профессиональную информацию, ваши анализы не будут столь ценными и эффективными с неактуальной информацией.
C. Если вы собираете не относящуюся к делу информацию, вы тратите впустую не только деньги, но и время. Ваши анализы не будут столь ценными и эффективными с неактуальной информацией.

5. **Accuracy отражает то на сколько получаемое из источника значение…**

A. Велико
B. Качественно
C. Точно
D. Разнообразно

Тема 6. Тест: разбалансированные датасеты и методы балансировки (тест считается сданным при верном ответе на 4 из 5 вопросов).

1. **Что такое разбалансированный датасет?**

A. Это датасет, в котором нет жёстко установленной формы
B. Это датасет, в котором один или несколько классов количественно преобладают над другими
C. Это датасет, в котором количество классов меньше или больше, чем установленная входная размерность нейронной сети
D. Это датасет, в котором классов больше, чем количество фичей

2. **Почему несбалансированность данных плохо влияет на обучение нейронных сетей?**

A. Алгоритм игнорирует малочисленный класс, что приводит плохому результату классификации
B. Алгоритм упирается во внутренние ошибки и не может выполнять задачу классификации
C. Алгоритм начинает слишком выделять малочисленный класс, для баланса его с более многочисленными классами, что приводит к нереалистичным результатам в продакшн
D. Алгоритм начинает плохо ориентироваться в форме данных, что приводит к ошибочной классификации

3. **Какие способы борьбы с разбалансировкой дата сета существуют?**

A. Сложение данных
B. Использовать дополнительные метрики качества
C. Добавить систему штрафов
D. Использовать разные алгоритмы
E. Перемешать данные

4. **По какому принципу работает алгоритм Undersampling?**

A. Алгоритм удаляет случайные элементы из большего класса
B. Алгоритм удаляет элементы с установленным шагом (Например, каждый третий)

C. Алгоритм удаляет каждый элемент, который выходит за рамки распределения
D. Алгоритм удаляет каждый элемент пары из большего набора

5. По какому принципу работает алгоритм Oversampling?
6.

A. Мы создаём элементы в непосредственной близости от уже существующих в меньшем наборе
B. Мы создаём новые элементы в виде среднего значения по всем данным
C. Мы создаём новые элементы, случайно выбирая их из установленного диапазона
D. Мы создаём новые элементы, выбирая их из диапазона нормального распределения случайным образом

Тема 7. Тест: библиотека Beautifulsoup. Парсинг данных из html страниц (тест считается сданным при верном ответе на 4 из 5 вопросов).

1. Что такое парсинг?

A. Это способ обработки данных на выходе из нейронной сети. Это делается для того, чтобы потом человек мог прочитать данные, которые ведёт модель
B. Это способ обработки баз данных для подачи на обучение нейронной сети
C. Это синтаксический анализ, разбор текста в синтаксическое дерево в соответствии с формальной грамматикой
D. Это способ анализа обработанных данных, позволяющий привести данные к неструктурированному виду

2. Что такое Веб-скрапинг?

A. Загрузка веб-страницы и попытка достать из неё информацию, обычно из формы, не предназначенной для этого, и в обход API и ограничений, а часто и правил пользования сайтом
B. Загрузка веб-страницы и попытка достать из неё информацию, обычно из формы, предназначенной для этого, и в обход API и ограничений, а часто и правил пользования сайтом
C. Загрузка веб-страницы и попытка достать из неё информацию, обычно из формы, не предназначенной для этого, но не в обход API и ограничений, и правил пользования сайтом
D. Загрузка веб-страницы и попытка достать из неё информацию, обычно из формы, предназначенной для этого, но не в обход API и ограничений, и правил пользования сайтом

3. Что такое краулинг?

A. Процесс обнаружения и сбора поисковым роботом (краулером) новых и обновленных страниц для добавления в индекс поисковых систем
B. Процесс обнаружения и сбора поисковым роботом (краулером) старых страниц для добавления в список для удаления
C. Процесс обнаружения и сбора поисковым роботом (краулером) новых и обновленных страниц для добавления их в список для проверки на нецензурную информацию
D. Процесс обнаружения и сбора поисковым роботом (краулером) старых страниц для добавления в список для дальнейшего обновления на них информации

4. Стоит ли пробовать парсить сайт, который не “поддаётся” путём отправки множества разнообразных запросов?

A. Да, это совершенно нормально
B. Да, но перед этим нужно убедиться, что вы не создадите сильную нагрузку на сайт
C. Нет, это совершено не этично

5. Если у нас есть доступный API, что это значит?

A. Будет проще разработать парсер
B. Есть возможность использовать API для частичной замены кода в парсере
C. Мы можем не писать свой парсер
D. Ни один ответ не является правильным

Тема 8. Тест: обработка категориальных признаков. LabelEncoder, One Hot encoding (тест считается сданным при верном ответе на 4 из 5 вопросов).

1. Что такое категориальные признаки?

A. Это признаки, которые имеют фиксированное значение
B. Это признаки, значения которых обозначают принадлежность объекта к какой-то категории
C. Это признаки, значения которых обозначают есть ли в выборке категории, которые обозначают другие признаки
D. Верных вариантов ответа нет

2. Есть ли универсальное решение как автоматически найти все категориальные признаки

A. Да, подобное решение имеется в библиотеке sklearn
B. Да, такое решение есть в библиотеке Tensorflow
C. Нет, таких решений не существует

3. Можно ли создавать новые категориальные признаки?

A. Да, это делается с помощью имеющихся категориальных признаков, путем манипуляции с ними
B. Нет, такой возможности не существует. Это противоречит природе данных

4. В чём особенность номинального атрибута?

A. Он зависит от ранга
B. Он зависит от своей позиции
C. Он зависит от своего ранга и позиции
D. Он не зависит от ранга
E. Он не зависит от позиции
F. Он не зависит от ранга и позиции

5. Выберите верный пример порядкового атрибута.

A. Состояние погоды

B. Размер футболок
C. Национальность человека
D. Тип деталей для строительства

Тема 9. Устный опрос: полная и условная вероятность, теорема Байеса.
Тема 10. Устный опрос: Байесовский вероятностный классификатор.
Тема 11. Устный опрос: метрики классификации. Матрица ошибок (Confusion -matrix) точность (Precision), полнота (recall), f1. ROC-AUC.
Тема 12. Устный опрос: кросс-валидация. Особенности применения.
Тема 13. Устный опрос: метод ближайших соседей k-NN. Метрики подсчета расстояния.
Тема 14. Устный опрос: метод опорных векторов (SVM). Линейно разделимые и неразделимые выборки, методы обработки. Плюсы и минусы алгоритма.
Тема 15. Устный опрос: линейная регрессия. Логистическая регрессия.
Тема 16. Устный опрос: метод наименьших квадратов. Средняя квадратичная ошибка, средняя абсолютная ошибка.
Тема 17. Устный опрос: решающие деревья (Decision tree).
Тема 18. Устный опрос: случайный лес (Random forest).
Тема 19. Устный опрос: термины, задачи и признаки машинного обучения.

5. Условия реализации ДПП

5.1. Организационные условия реализации ДПП

Наименование аудитории	Вид занятия	Наименование оборудования, программного обеспечения
Компьютерный класс	Лекции	**Материальное обеспечение:** компьютер, мультимедийный проектор, экран, доска, пишущий инструмент, **Программное обеспечение:** Anaconda
Компьютерный класс	Практические занятия	**Материальное обеспечение:** компьютер, мультимедийный проектор, экран, доска, пишущий инструмент, **Программное обеспечение:** Anaconda
Компьютерный класс	Самостоятельная работа	**Материальное обеспечение:** компьютер, мультимедийный проектор, экран, доска, пишущий инструмент, **Программное обеспечение:** Anaconda
Компьютерный класс	Итоговая аттестация	**Материальное обеспечение:** компьютер, мультимедийный проектор, экран, доска, пишущий инструмент, **Программное обеспечение:** Anaconda

5.2. Педагогические условия реализации ДПП

Реализация программы обеспечивается преподавательским составом, удовлетворяющим следующим условиям:
- наличие высшего образования, соответствующее профилю программы, из числа штатных преподавателей, или привлеченных на условиях почасовой оплаты труда;
- опыт практической деятельности в соответствующей сфере из числа штатных преподавателей или привлеченных на условиях почасовой оплаты труда.

5.3. Учебно-методическое обеспечение ДПП

1. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных / Силен Д., Мейсман А., Али М.; пер. с англ. Матвеев Е. – СПб.: Питер, 2020. – 334 с. (https://library.bmstu.ru/Catalog/Details/544371)
2. Шелухин О.И., Ерохин С.Д., Полковников М.В. Технологии машинного обучения в сетевой безопасности / Шелухин О.И., Ерохин С.Д., Полковников М.В.; ред. Шелухин О.И. – М.: Горячая линия-Телеком, 2021. – 359 с. (https://library.bmstu.ru/Catalog/Details/555230)
3. Сузи Р. Python. В подлиннике: Наиболее полное руководство / Сузи Р. – СПб.: БХВ-Петербург, 2002. – 747 с. (https://library.bmstu.ru/Catalog/Details/84709)
4. Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение / Плас Дж. Вандер; пер. с англ. Пальти И. – СПб.: Питер, 2020. – 572 с. (https://library.bmstu.ru/Catalog/Details/550326)
5. Круз Р.Л. Структуры данных и проектирование программ: [учеб. пособие] / Круз Р.Л.; пер. 3-го англ. изд. Финогенов К.Г. – М.: БИНОМ. Лаборатория знаний, 2017. – 765 с. (https://library.bmstu.ru/Catalog/Details/476469)
6. Галкин С.В., Панов В.Ф., Петрухина О.С. Краткий курс теории вероятностей: учеб. пособие / Галкин С.В., Панов В.Ф., Петрухина О.С.; МГТУ им. Н.Э. Баумана. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2007. – 54 с. (https://library.bmstu.ru/Catalog/Details/178041)
7. Хливненко Л.В., Пятакович Ф.А. Практика нейросетевого моделирования: учебное пособие / Хливненко Л.В., Пятакович Ф.А. – 2-е изд., стер. – СПб.: Лань, 2021. – 196 с. (https://library.bmstu.ru/Catalog/Details/556123)
8. Грановская Р.М., Березная И.Я. Интуиция и искусственный интеллект / Грановская Р.М., Березная И.Я.; Ленинградский гос. ун-т. – Л.: Изд-во Ленинградского ун-та, 1991. – 268 с. (https://library.bmstu.ru/Catalog/Details/503904)
9. Шумский С.А. Машинный интеллект. Очерки по теории машинного обучения и искусственного интеллекта / Шумский С.А.; Московский физико-технический ин-т (национальный исследовательский ун-т). – М.: РИОР: Инфра-М, 2021. – 339 с. (https://library.bmstu.ru/Catalog/Details/554838)
10. Берикашвили В.Ш., Оськин С.П. Статистическая обработка данных, планирование эксперимента и случайные процессы: учебное пособие для вузов / Берикашвили В.Ш., Оськин С.П. - 2-е изд., испр. и доп. – М.: Юрайт, 2021. – 163 с. (https://library.bmstu.ru/Catalog/Details/555713)
11. Вьюгин В.В. Элементы математической теории машинного обучения: учеб. пособие для вузов / Вьюгин В.В.; Моск. физико-техн. ин-т (гос. ун-т), РАН. Ин-т проблем передачи информации им. А.А. Харкевича. – М.: МФТИ - ИППИ РАН, 2010. – 231 с. (https://library.bmstu.ru/Catalog/Details/229808)
12. Меженная Н.М. Основы теории вероятностей и математической статистики: курс лекций / Меженная Н.М.; МГТУ им. Н.Э. Баумана. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2016. – 108 с. (https://library.bmstu.ru/Catalog/Details/465260)
13. Амосов А.А., Дубинский Ю.А., Копченова Н.В. Вычислительные методы: учеб. пособие / Амосов А.А., Дубинский Ю.А., Копченова Н.В. - 4-е изд., стер. – СПб.: Лань, 2014. – 671 с. (https://library.bmstu.ru/Catalog/Details/379247)
14. Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознования образов и анализа разнотипной информации / Лбов Г.С., Бериков В.Б. – Новосибирск: Изд-во Ин-та математики, 2005. – 217 с. (https://library.bmstu.ru/Catalog/Details/126725)
15. Рассел С., Норвиг П. Искусственный интеллект. Современный подход / Рассел С., Норвиг П.; пер. с англ. и ред. Птицын К.А. - 2-е изд. – М.: Изд. дом «Вильямс», 2018. – 1407 с. (https://library.bmstu.ru/Catalog/Details/518925)

5.4. Методические рекомендации

ДПП построена по тематическому принципу, каждый раздел представляет собой логически завершенный материал.

Преподавание программы основано на личностно-ориентированной технологии образования, сочетающей два равноправных аспекта этого процесса: обучение и учение. Личностно-ориентированный подход развивается при участии слушателей в активной работе на практических занятиях. Личностно-ориентированный подход направлен, в первую очередь, на развитие индивидуальных способностей обучающихся, создание условий для развития творческой активности слушателя и разработке инновационных идей, а также на развитие самостоятельности мышления при решении учебных задач разными способами, нахождение рационального варианта решения, сравнения и оценки нескольких вариантов их решения и т.п. Это способствует формированию приемов умственной деятельности по восприятию новой информации, ее запоминанию и осознанию, созданию образов для сложных понятий и процессов, приобретению навыков поиска решений в условиях неопределенности.

Лекции проводятся для приобретения навыков реализации знаний в предметной области, с использованием активных методов обучения.

Практические занятия проводятся для приобретения навыков решения практических задач в предметной области. Задания, выполняемые на практических занятиях, выполняются с использованием активных и интерактивных методов обучения.

Самостоятельная работа слушателей предназначена для проработки дополнительных источников информации. При изучении ДПП предусмотрены следующие методы организации и осуществления учебно-познавательной деятельности:

- объяснительно-иллюстративный метод;
- репродуктивный метод;
- частично-поисковый метод.

При изучении ДПП предусмотрены активные формы проведения занятий:

– управляемая дискуссия;
– разбор конкретных ситуаций.

6. Формы итоговой аттестации ДПП

Итоговая аттестация проводится в форме зачета для проверки сформированности компетенций, полученных в рамках ДПП.

Зачет проводится в формате тестирования. Результатом зачета служат правильные ответы на вопросы билета.

По результатам итоговой аттестации обучающемуся выставляется оценка «ЗАЧТЕНО/НЕ ЗАЧТЕНО»:

Оценка «ЗАЧТЕНО» выставляется обучающемуся, который:

- ответил на 8 из 12 вопросов теста;
- продемонстрировал необходимые систематизированные знания и достаточную степень владения принципами предметной области программы, понимание их особенностей и взаимосвязь между ними в течение всего срока обучения по ДПП.

Оценка «НЕ ЗАЧТЕНО» ставятся обучающемуся, который:

- ответил менее, чем на 8 из 12 вопросов теста;
- имеет крайне слабые теоретические и практические знания, обнаруживает неспособность к построению самостоятельных заключений.

7. Оценочные материалы итоговой аттестации

7.1. Паспорт комплекта оценочных средств

Предметы оценивания	Объекты оценивания	Показатели оценки

ПК-1. Способен создавать информационные системы, понимание существующих подходов к верификации моделей программного обеспечения	Ответы на вопросы	Количество правильных ответов
ОПК-5. Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем	Ответы на вопросы	Количество правильных ответов

7.2. Комплект оценочных средств

7.2.1. Темы для подготовки к зачету:

1. Метод главных компонент.
2. Кластеризация данных.
3. Подготовка датасетов.
4. Методы балансировки датасетов.
5. Парсинг данных.
6. Полная и условная вероятность.
7. Метрики классификации.
8. Виды регрессии.
9. Решающие деревья и случайный лес.
10. Ансамбли алгоритмов.

7.2.2. Пример билета:

1 – Алгоритм машинного обучения для визуализации, разработанный Лоренсом ван дер Маатеном и Джеффри Хинтоном, является техникой нелинейного снижения размерности, хорошо подходящей для вложения данных высокой размерности для визуализации в пространство низкой размерности это …
A – SNS
B – t-SNE
C – SNE
D – t-SNS

2 – Алгоритм обучения без учителя. Смысл алгоритма заключается в наблюдении за набором немаркированных данных для автоматического обнаружения скрытой структуры, а также для обнаружения закономерности в немаркированных данных это…
A – s-means
B – d-means
C – c-means
D – k-means

3 – Алгоритм позволяет разбить имеющееся множество элементов мощностью на заданное число нечётких множеств. Метод нечеткой кластеризации можно рассматривать как усовершенствованный метод *k*-средних, при котором для каждого элемента из рассматриваемого множества рассчитывается степень его принадлежности каждому из кластеров

A – s-means
B – d-means
C – c-means
D – k-means

4 – Это общее семейство алгоритмов кластеризации, которые создают вложенные кластеры путем их последовательного слияния или разделения. Всё представляется в виде дерева, где корень - уникальный кластер, который собирает все образцы, а листья – это кластеры только с одним образцом. Что из вариантов ниже подходит под описание?
A – Иерархическая кластеризация
B – Решающие деревья
C – Спектральная кластеризация
D – Семейство методов сдвига

5 – Алгоритм кластеризации, основанный на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены, помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью. Какой вариант подходит под описание?
A – SCAN_SCP
B – BSCAN_V2
C – DBSCAN
D – BdD_SCAN

6 – Как назвать обработанную и структурированную информацию в табличном виде, если строки таблицы — это объекты, а столбцы признаки?
A – Массив
B – Таблица
C – Датасет
D – Словарь

7 – Если стоит задача определения типа покрытия пола с помощью машинного обучения и есть набор данных из разных изображений пола. Как правильно будет разделить эти данные?
A – Разделить фото по этим типам
B – Разделить фото по разрешению
C – Разделить фото в зависимости от используемой цветовой палитры
D – Разделить фото на случайные, но равные между собой части

8 – Что имеется ввиду под разбалансированным датасетом?
A – Когда в датасете большая количественная разница между примерами классов, которые нужно предсказывать
B – Когда датасет имеет разное количество признаков для разных объектов
C – Когда в датасете количество признаков не равно количеству признаков

9 – Что из списка подразумевает такой способ работы: Предположим, что некоторый признак может принимать 10 разных значений. Значит создаётся 10 признаков, все из которых равны нулю за исключением одного. На позицию, соответствующую численному значению признака мы помещаем 1
A – LabelEncoder
B – One-Hot Encoding
C – Hashing trick

10 – Для каких целей служит библиотека beautifulsoup?
A – Для создания программ, оптимизирующих процесс разработки ИИ решений
B – Для автоматической разметки очень больших датасетов
C – Для изучения и анализа не структурированных данных, что позволяет не тратить время на их структурирование
D – Для парсинга HTML и XML документов. Часто используется для скрапинга веб-страниц

11 – Какое описание подходит методу ближайших соседей?
A – Простейший метрический классификатор, основанный на оценивании сходства объектов
B – Сложный в своей реализации метрический классификатор, основанный на оценивании сходства объектов
C – Простейший метрический классификатор, основанный на теории дублированных объектов в зависимости от их положения в графе важности для данной задачи

12 – Метод машинного обучения целью которого является попытка классифицировать входные наборы данных в один из двух классов это…
A – Метод опорных векторов
B – Метод байесовского распределения
C – Метод двойного распределения