Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Максимов Алексей Борисович
Должность: директор департамента по образовательной политике
Дата подписания: 31.05.2024 17:12:15
Уникальный программный ключ:
8db180d1a3f02ac9e60521a5672742735c18b1d6

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение
высшего образования
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
Факультет информационных технологий

УТВЕРЖДАЮ
Декан факультета
«Информационные технологии»
_____ / Д.Г.Демидов /
«15» февраля 2024г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Хранилища данных и технологий Big data»

Направление подготовки/специальность
01.04.02 Прикладная математика и информатика

Профиль/специализация
«Системная аналитика больших данных»

Квалификация
магистр

Формы обучения
очная

Москва, 2024 г.

Разработчик(и):

к.ф-м.н., доцент /А.В. Осипов /

Согласовано:
Заведующий кафедрой «Прикладная информатика»,
к.э.н., доцент /С.В. Суворов /

Содержание

Оглавление

1 Цели, задачи и планируемые результаты обучения по дисциплине 4
2 Место дисциплины в структуре образовательной программы 5
3 Структура и содержание дисциплины 5
3.1 Виды учебной работы и трудоемкость 5
3.2 Тематический план изучения дисциплины 6
3.3 Содержание дисциплины 6
3.4 Тематика семинарских/практических и лабораторных занятий 7
3.5 Тематика курсовых проектов (курсовых работ) 7
4 Учебно-методическое и информационное обеспечение 8
4.1 Нормативные документы и ГОСТы 8
4.2 Основная литература 8
4.3 Дополнительная литература 8
4.4 Электронные образовательные ресурсы 8
4.5 Лицензионное и свободно распространяемое программное обеспечение 8
4.6 Современные профессиональные базы данных и информационные справочные системы 8
5 Материально-техническое обеспечение 8
6 Методические рекомендации 9
6.1 Методические рекомендации для преподавателя по организации обучения 9
6.2 Методические указания для обучающихся по освоению дисциплины 9
7 Фонд оценочных средств 9
7.1 Методы контроля и оценивания результатов обучения 9
7.2 Шкала и критерии оценивания результатов обучения 10
7.3 Оценочные средства 12

1 Цели, задачи и планируемые результаты обучения по дисциплине

К **основным целям** освоения дисциплины «Хранилища данных и технологии Big Data» следует отнести:

– расширенное формирование у студентов представления о принципах и методах машинного обучения;

– знакомство студентов с современными методами работы с большими данными.

К **основным задачам** освоения дисциплины «Хранилища данных и технологии Big Data» следует отнести:

– освоение методологии обработки больших данных;

– использование компьютерных технологий реализации методов машинного обучения..

Обучение по дисциплине «Хранилища данных и технологий Big data» направлено на формирование у обучающихся следующих компетенций:

Код и наименование компетенций	Индикаторы достижения компетенции
УК-2. Способен управлять проектом на всех этапах его жизненного цикла	**знать:** этапы жизненного цикла проекта; этапы разработки и реализации проекта; методы разработки и управления проектами; **уметь:** разработать проект с учетом анализа альтернативных вариантов его реализации, определять целевые этапы, основные направления работ; объяснить цели и сформулировать задачи, связанные с подготовкой и реализацией проекта; управлять проектом на всех этапах его жизненного цикла; **владеть:** методиками разработки и управления проектом; методами оценки потребности в ресурсах и эффективности проекта.
ОПК-4. Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности с учетом требований информационной безопасности	**знать:** современные информационно-коммуникационные и интеллектуальные технологии, инструментальные среды, программно-технические платформы для решения профессиональных задач с учетом требований информационной безопасности; **уметь:** обосновывать выбор современных информационно-коммуникационных и интеллектуальных технологий, разрабатывать оригинальные программные средства для решения профессиональных задач с учетом требований информационной безопасности; **владеть:** навыками разработки оригинальных программных средств, в том числе с использованием современных информационно-коммуникационных и интеллектуальных технологий, для

	решения профессиональных задач с учетом требований информационной безопасности.

2 Место дисциплины в структуре образовательной программы

Дисциплина «Хранилища данных и технологии Big Data» относится к базовой части основной образовательной программы магистратуры.

Она взаимосвязана логически и содержательно - методически со следующими дисциплинами ООП:

- Современные проблемы прикладной математики и информатики;
- Математическое моделирование открытых данных;
- Непрерывные математические модели.

3 Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 4 зачетных единицы т.е. 144 академических часа 44 часов лабораторные занятия и 100 часов – самостоятельная работа студентов, зачет - 2 семестр и экзамен - 3 семестр.

3.1 Виды учебной работы и трудоемкость (по формам обучения)

3.1.1 Очная форма обучения

№ п/п	Вид учебной работы	Количество часов	Семестры	
			2	3
1	**Аудиторные занятия**	**44**		
	В том числе:			
1.1	Лекции			
1.2	Семинарские/практические занятия			
1.3	Лабораторные занятия		24	20
2	**Самостоятельная работа**	**100**		
	В том числе:			
2.1	Подготовка к практическим занятиям		48	52
3	**Промежуточная аттестация**			
	Зачет/диф.зачет/экзамен		зачет	экзамен
	Итого:	**144**	72	72

3.2 Тематический план изучения дисциплины (по формам обучения)

3.2.1 Очная форма обучения

№ п/п	Разделы/темы дисциплины	Трудоемкость, час					
		Всего	Аудиторная работа				Самостоятельная работа
			Лекции	Семинарские/ практические занятия	Лабораторные занятия	Практическая подготовка	
1.1	Тема 1. Область применения хранилищ данных и больших данных и их преимущества	36			12		24
1.2	Тема 2. Виды данных	36			12		24
1.3	Тема 3. Экосистема больших данных	36			10		26
1.4	Тема 4. Технологии Data Mining	36			10		26
Итого		**144**			**36**		**100**

3.3 Содержание дисциплины

Второй семестр

1. Область применения хранилищ данных и больших данных и их преимущества
2. Виды данных
2.1 Структурированные данные
2.2 Неструктурированные данные
2.3 Данные на естественном языке
2.4 Машинные данные
2.5 Графовые, или сетевые, данные
2.6 Аудио, видео и графика
2.7 Потоковые данные
3. Экосистема больших данных
3.1 Распределенные файловые системы
3.2 Инфраструктура распределенного программирования
3.3 Инфраструктура интеграции данных
3.4 Инфраструктуры машинного обучения
3.5 Базы данных NoSQL

Третий семестр

4. Технологии Data Mining
4.1 Задачи Data Mining

Обзор задач Data Mining. Стандартизация подхода к решению задач Data Mining. Процесс CRISP-DM. Виды данных. Кластеризация, классификация, регрессия. Понятие модели и алгоритма обучения.

4.2 Задача кластеризации и EM-алгоритм

Постановка задачи кластеризации. Функции расстояния. Критерии качества кластеризации. EM-алгоритм. K-means и модификации.

4.3 Различные алгоритмы кластеризации

Иерархическая кластеризация. Agglomerative и Divisive алгоритмы. Различные виды расстояний между кластерами. Stepwise-optimal алгоритм. Случай неэвклидовых пространств. Критерии выбора количества кластеров: rand, silhouette. DBSCAN.

4.4 Задача классификации

Постановка задач классификации и регрессии. Теория принятия решений. Виды моделей. Примеры функций потерь. Переобучение. Метрики качества классификации. MDL. Решающие деревья. Алгоритм CART.

4.5 Обработка текстов, Naive Bayes

Условная вероятность и теорема Байеса. Нормальное распределение. Naive Bayes: multinomial, binomial, gaussian. Сглаживание. Генеративная модель NB и байесовский вывод. Графические модели.

4.6 Линейные модели для классификации и регрессии

Обобщенные линейные модели. Постановка задачи оптимизации. Примеры критериев. Градиентный спуск. Регуляризация. Метод Maximum Likelihood. Логистическая регрессия.

4.7 Машина опорных векторов

Разделяющая поверхность с максимальным зазором. Формулировка задачи оптимизации для случаев линейно-разделимых и линейно-неразделимых классов. Сопряженная задача. Опорные векторы. KKT-условия. SVM для задач классификации и регрессии. Kernel trick. Теорема Мерсера. Примеры функций ядра.

4.8 Методы снижения размерности пространства

Проблема проклятия размерности. Отбор и выделение признаков. Методы выделения признаков (feature extraction). Метод главных компонент (PCA). Метод независимых компонент (ICA). Методы основанные на автоэнкодерах. Методы отбора признаков (feature selection). Методы основанные на взаимной корреляции признаков. Метод максимальной релевантность и минимальной избыточности (mRMR). Методы основанные на деревьях решений.

4.9 Алгоритмические композиции

Комбинации классификаторов. Модельные деревья решений. Смесь экспертов. Stacking. Стохастические методы построения ансамблей классификаторов. Bagging. RSM. Алгоритм RandomForest.

4.10 Глубокие нейронные сети

Трудности обучения многослойного персептрона. Предобучение используя РБМ. Глубокий автоэнкодер, глубокая многослойная нейросеть. Deep belief network и deep Boltzmann machine. Сверточные сети.

3.4 Тематика семинарских/практических и лабораторных занятий

Тема 1. Область применения хранилищ данных и больших данных и их преимущества

Тема 2. Виды данных

Тема 3. Экосистема больших данных

Тема 4. Технологии Data Mining.

3.5 Тематика курсовых проектов (курсовых работ)

Не предусмотрено

4 Учебно-методическое и информационное обеспечение

4.1 Нормативные документы и ГОСТы

1. Федеральный государственный образовательный стандарт высшего образования (уровень магистратуры) по направлению подготовки 01.04.02 Прикладная математика и информатика, утвержденный приказом Министерства науки и высшего образования Российской Федерации от 10.01.2018 №13.

4.2 Основная литература

1. Смирнов, М. В. Проектирование и администрирование хранилищ и баз данных : методические рекомендации / М. В. Смирнов, Р. С. Толмасов. — Москва : РТУ МИРЭА, 2022. — 33 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/310871 (дата обращения: 27.09.2023). — Режим доступа: для авториз. пользователей.

2. Орешков, В. И. Хранилища данных и OLAP-технологии : учебное пособие / В. И. Орешков. — Рязань : РГРТУ, 2017. — 64 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/167981 (дата обращения: 27.09.2023). — Режим доступа: для авториз. пользователей.

4.3 Дополнительная литература

1. Точилкина, Т. Е. Хранилища данных и средства бизнес-аналитики : учебное пособие / Т. Е. Точилкина, А. А. Громова. — Москва : Финансовый университет, 2017. — 161 с. — ISBN 978-5-7942-1387-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/208367 (дата обращения: 27.09.2023). — Режим доступа: для авториз. пользователей.

2. Смирнов, М. В. Проектирование и администрирование хранилищ и баз данных : методические рекомендации / М. В. Смирнов, Р. С. Толмасов. — Москва : РТУ МИРЭА, 2022. — 33 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/310871 (дата обращения: 27.09.2023). — Режим доступа: для авториз. пользователей.

4.4 Электронные образовательные ресурсы

1. https://online.mospolytech.ru/course/view.php?id=2241 - Хранилища данных

4.5 Лицензионное и свободно распространяемое программное обеспечение

1. Microsoft Windows.
2. Microsoft Visio.
3. Microsoft Office.

4.6 Современные профессиональные базы данных и информационные справочные системы

Не предусмотрено

5 Материально-техническое обеспечение

Четыре компьютерных класса Ауд. АВ4805, АВ4809, АВ4810, АВ4811, оснащенные методическими материалами по дисциплине (лекции, практические задания).

6 Методические рекомендации

6.1 Методические рекомендации для преподавателя по организации обучения

1. При подготовке к занятиям следует предварительно проработать материал занятия, предусмотрев его подачу точно в отведенное для этого время занятия. Следует подготовить необходимые материалы – теоретические сведения, задачи и др. При проведении занятия следует контролировать подачу материала и решение заданий с учетом учебного времени, отведенного для занятия.

2. При проверке работ и отчетов следует учитывать не только правильность выполнения заданий, но и оптимальность выбранных методов решения, правильность выполнения всех его шагов.

3. При организации и проведения зачетов в практико-ориентированной форме следует использовать утвержденные кафедрой методические рекомендации..

6.2 Методические указания для обучающихся по освоению дисциплины

Изучение дисциплины осуществляется в строгом соответствии с целевой установкой в тесной взаимосвязи учебным планом. Основой теоретической подготовки студентов являются аудиторные занятия, лекции, лабораторные работы.

В процессе самостоятельной работы студенты закрепляют и углубляют знания, полученные во время аудиторных занятий, дорабатывают конспекты и записи, готовятся к проведению и обрабатывают результаты лабораторных работ, готовятся к промежуточной аттестации, а также самостоятельно изучают отдельные темы учебной программы.

На занятиях студентов, в том числе предполагающих практическую деятельность, осуществляется закрепление полученных, в том числе и в процессе самостоятельной работы, знаний. Особое внимание обращается на развитие умений и навыков установления связи положений теории с профессиональной деятельностью будущего специалиста.

Самостоятельная работа осуществляется индивидуально. Контроль самостоятельной работы организуется в двух формах:

- *самоконтроль и самооценка студента;*
- *контроль со стороны преподавателей (текущий и промежуточный).*

Критериями оценки результатов самостоятельной работы студента являются:

- *уровень освоения студентом учебного материала;*
- *умения студента использовать теоретические знания при выполнении практических задач;*
- *сформированность компетенций;*

оформление материала в соответствии с требованиями.

.

7 Фонд оценочных средств

7.1 Методы контроля и оценивания результатов обучения

В процессе обучения используются следующие оценочные формы самостоятельной работы студентов, оценочные средства текущего контроля успеваемости и промежуточных аттестаций:

- выполнение лабораторных работ;
- зачет;
- экзамен.

Методика преподавания дисциплины «Хранилища данных и технологии Big Data» и реализация компетентного подхода в изложении и восприятии материала предусматривает использование следующих активных и интерактивных форм проведения групповых, индивидуальных, аудиторных занятий в сочетании с внеаудиторной работой с целью формирования и развития профессиональных навыков обучающихся:

– подготовка к выполнению практических работ в компьютерных классах вуза;

– подготовка, представление и обсуждение презентаций на семинарских занятиях;

– организация и проведение интерактивных форм текущего контроля знаний студентов в форме выполнения индивидуальных заданий.

Удельный вес занятий, проводимых в интерактивных формах, определен главной целью образовательной программы, особенностью контингента обучающихся и содержанием дисциплины «Хранилища данных и технологии Big Data» и в целом по дисциплине составляет 50% аудиторных занятий. Занятия лекционного типа составляют 50% от объема аудиторных занятий.

7.2 Шкала и критерии оценивания результатов обучения

Показателем оценивания компетенций на различных этапах их формирования является достижение обучающимися планируемых результатов обучения по дисциплине.

Показатель	**Критерии оценивания**			
	2	**3**	**4**	**5**
УК-2. Способен управлять проектом на всех этапах его жизненного цикла				
УК-1.1. Знать: этапы жизненного цикла проекта; этапы разработки и реализации проекта; методы разработки и управления проектами УК-1.2. Уметь: разработать проект с учетом анализа альтернативных вариантов его реализации, определять целевые этапы, основные направления работ; объяснить цели и сформулировать	Обучающийся демонстрирует полное отсутствие или недостаточное соответствие материалу дисциплины знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3).	Обучающийся демонстрирует неполное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3). Допускаются значительные ошибки, проявляется недостаточность знаний, по ряду показателей, обучающийся испытывает значительные затруднения	Обучающийся демонстрирует частичное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3). Но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях.	Обучающийся демонстрирует полное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3). Свободно оперирует приобретенными знаниями.

задачи, связанные с подготовкой и реализацией проекта; управлять проектом на всех этапах его жизненного цикла УК-2.3. Владеть: методиками разработки и управления проектом; методами оценки потребности в ресурсах и эффективности проекта		при оперировании знаниями при их переносе на новые ситуации.		
ОПК-4. Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности с учетом требований информационной безопасности				
ОПК-4.1. Знать: современные информационно-коммуникационные и интеллектуальные технологии, инструментальные среды, программно-технические платформы для решения профессиональных задач с учетом требований информационной безопасности; ОПК-4.2. Уметь: обосновывать выбор современных информационно-коммуникационных и интеллектуальных технологий,	Обучающийся демонстрирует полное отсутствие или недостаточное соответствие материалу дисциплины знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3).	Обучающийся демонстрирует неполное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3). Допускаются значительные ошибки, проявляется недостаточность знаний, по ряду показателей, обучающийся испытывает значительные затруднения при оперировании знаниями при их переносе на	Обучающийся демонстрирует частичное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3). Но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях.	Обучающийся демонстрирует полное соответствие следующих знаний, указанных в индикаторах компетенций дисциплины «Знать» (см. п. 3). Свободно оперирует приобретенными знаниями.

разрабатывать оригинальные программные средства для решения профессиональных задач с учетом требований информационной безопасности ОПК-4.3. Владеть: навыками разработки оригинальных программных средств, в том числе с использованием современных информационно-коммуникационных и интеллектуальных технологий, для решения профессиональных задач с учетом требований информационной безопасности		новые ситуации.		

7.3 Оценочные средства

7.3.1 Текущий контроль

Зачетные вопросы

1. Область применения хранилищ данных и больших данных и их преимущества
2. Виды данных. Структурированные данные
3. Виды данных. Неструктурированные данные
4. Виды данных. Данные на естественном языке
5. Виды данных. Машинные данные
6. Виды данных. Графовые, или сетевые, данные
7. Виды данных. Аудио, видео и графика
8. Виды данных. Потоковые данные
9. Понятие гиперкуба. OLAP.
10. ROLAP-системы
11. MOLAP-системы
12. HOLAP-системы
13. Распределенные файловые системы

14. Базы данных NoSQL

Экзаменационные вопросы

1. Обзор задач Data Mining. Кластеризация, классификация, регрессия.
2. Постановка задачи кластеризации. Функции расстояния. Критерии качества кластеризации.
3. EM-алгоритм.
4. K-means и модификации.
5. Иерархическая кластеризация. Agglomerative и Divisive алгоритмы.
6. Постановка задач классификации и регрессии. Теория принятия решений.
7. Виды моделей. Примеры функций потерь. Переобучение. Метрики качества классификации.
8. MDL. Решающие деревья.
9. Алгоритм CART.
10. Условная вероятность и теорема Байеса. Нормальное распределение.
11. Naive Bayes: multinomial, binomial, gaussian. Сглаживание. Генеративная модель NB и байесовский вывод.
12. Обобщенные линейные модели. Постановка задачи оптимизации. Примеры критериев. Градиентный спуск.
13. Разделяющая поверхность с максимальным зазором. Формулировка задачи оптимизации для случаев линейно-разделимых и линейно-неразделимых классов.
14. KKT-условия. SVM для задач классификации и регрессии.
15. Проблема проклятия размерности. Отбор и выделение признаков. Методы выделения признаков (feature extraction).
16. Метод главных компонент (PCA).
17. Метод независимых компонент (ICA).
18. Методы, основанные на автоэнкодерах.
19. Комбинации классификаторов. Модельные деревья решений.
20. Алгоритм RandomForest.
21. Глубокий автоэнкодер, глубокая многослойная нейросеть.
22. Deep belief network и deep Boltzmann machine.
23. Сверточные сети.

7.3.2 Промежуточная аттестация

Промежуточная аттестация обучающихся в форме экзамена проводится по результатам выполнения всех видов учебной работы, предусмотренных учебным планом по данной дисциплине (модулю), при этом учитываются результаты текущего контроля успеваемости в течение семестра. Оценка степени достижения обучающимися планируемых результатов обучения по дисциплине (модулю) проводится преподавателем, ведущим занятия по дисциплине (модулю) методом экспертной оценки. По итогам промежуточной аттестации по дисциплине (модулю) выставляется оценка «отлично», «хорошо», «удовлетворительно» или «неудовлетворительно».

Шкала оценивания	***Описание***

Отлично	*Выполнены все виды учебной работы, предусмотренные учебным планом. Студент демонстрирует соответствие знаний, умений, навыков приведенным в таблицах показателей, оперирует приобретенными знаниями, умениями, навыками, применяет их в ситуациях повышенной сложности. При этом могут быть допущены незначительные ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации.*
Хорошо	*Выполнены все виды учебной работы, предусмотренные учебным планом. Студент демонстрирует неполное, правильное соответствие знаний, умений, навыков приведенным в таблицах показателей, либо если при этом были допущены 2-3 несущественные ошибки.*
Удовлетворительно	*Выполнены все виды учебной работы, предусмотренные учебным планом. Студент демонстрирует соответствие знаний, в котором освещена основная, наиболее важная часть материала, но при этом допущена одна значительная ошибка или неточность.*
Неудовлетворительно	*Не выполнен один или более видов учебной работы, предусмотренных учебным планом. Студент демонстрирует неполное соответствие знаний, умений, навыков приведенным в таблицах показателей, допускаются значительные ошибки, проявляется отсутствие знаний, умений, навыков по ряду показателей, студент испытывает значительные затруднения при оперировании знаниями и умениями при их переносе на новые ситуации.*

Экзаменационное задание выполняется студентом индивидуально, по итогам изучения дисциплины или ее части. При этом достижение порогового результата работы над экзаменационным заданием соответствует описанному в п. 3 данного документа этапу освоения соответствующих компетенций на базовом или продвинутом уровне.

Базовый уровень: способность выполнять полученное задание, применяя полученные знание и умения на практике, владеть соответствующими индикаторами компетенции при выполнении задания.

Продвинутый уровень: способность выполнять полученное задание и решать самостоятельно сформированные задачи, применяя полученные знание и умения на практике. Уверенно владеть соответствующими индикаторами компетенции при выполнении задания, комбинировать их между собой и с индикаторами других компетенций для достижения проектных результатов.