

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
(ВлГУ)



УТВЕРЖДАЮ
Проректор
по образовательной деятельности

А.А. Панфилов

« 19 » 06 2019 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
«Технологии анализа данных»

Направление подготовки: **09.04.04 «Программная инженерия»**

Профиль/программа подготовки: **Разработка программно-информационных систем**

Уровень высшего образования: **магистратура**

Форма обучения: **очная**

Семестр	Трудоём- кость зач. Ед./час.	Лекции, час.	Практич. Занятия, час.	Лаборат. Работы, час.	СРС, час.	Форма промежуточной аттестации (экз./зачет)
1	4/144	18		18	81	Экзамен – 27 ч.
Итого	4/144	18		18	81	Экзамен – 27 ч.

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целями освоения дисциплины «Технологии анализа данных» являются:

1. Ознакомление студентов с основными принципами машинного обучения, а именно:
 - подходами к предварительной обработке данных;
 - видами задач и методов машинного обучения (линейная регрессия, кластеризация, прогнозирование временных рядов).
2. Формирование у студентов практических навыков сбора, хранения, обработки данных и решения задач статистического анализа данных на языке R, прикладного анализа данных на языке Python.

Задачи дисциплины:

1. Повышение уровня компетенции студентов за счет приобретения соответствующих знаний и практических умений.
2. Формирование теоретических и методологических основ в области анализа данных, а также практических навыков решения задач статистического и прикладного анализа данных, машинного обучения, реализованных в специализированных программных продуктах.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

Дисциплина «Технологии анализа данных» относится к вариативной части учебного плана.

Пререквизиты дисциплины: «Технологии программирования», «Управление данными», «Методы анализа данных»

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП

Код формируемых компетенций	Уровень освоения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций (показатели освоения компетенции)
1	2	3
ОПК-1	Частичное освоение	Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности. Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-

		экономических и профессиональных знаний. Иметь навыки: теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте.
ОПК-7	Частичное освоение	Знать: методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях. Уметь: применять методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях. Иметь навыки: применения методов и средств получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях.
ПК-1	Частичное освоение	Знать: Актуальную нормативную документацию в соответствующей области знаний; Методы анализа научных данных; Методы и средства планирования и организации исследований и разработок Уметь: Применять актуальную нормативную документацию в соответствующей области знаний; Оформлять результаты научно-исследовательских и опытно-конструкторских работ Иметь навыки: Осуществления разработки планов и методических программ проведения исследований и разработок; Организации сбора и изучения научно-технической информации по теме исследований и разработок; Проведения анализа научных

		данных, результатов экспериментов и наблюдений; Осуществления теоретического обобщения научных данных, результатов экспериментов и наблюдений
--	--	--

4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины составляет 4 зачетных единиц, 144 часа.

№ п/п	Раздел (тема) дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоемкость (в часах)					Объем учебной работы, с применением интерактивных методов (часы/%)	Формы текущего контроля успеваемости, форма промежуточной аттестации	
				Лекции	Практические занятия	Лабораторные работы	СРС	КП / КР			
1	Введение в технологии обработки и анализа данных	1	1-2	4		2	10		3/50		
2	Проектирование систем анализа и обработки данных	1	3-4	2		2	10		3/75		
3	Технология OLAP	1	5-6	2		2	10		3/75	РК 1	
4	Понятие ETL. Разработка ETL-процесса с помощью инструмента Pentaho Data Integration	1	7-8	2		2	10		3/75		
5	Язык статистического анализа данных R. Интегрированная среда разработки RStudio	1	9-10	2		2	10		3/75		
6	Методы машинного обучения: линейная регрессия. Прикладной анализ данных с помощью Python.	1	11-12	2		2	10		3/75	РК 2	
7	Методы машинного обучения: кластеризация. Прикладной анализ данных с помощью Python.	1	13-15	2		2	10		4/100		
8	Методы машинного обучения: прогнозирование временных рядов. Прикладной анализ данных с помощью Python.	1	16-18	2		4	11		4/67	РК 3	
Итого за 1 семестр						18		18	81		Экзамен
Всего						18		18	81		Экзамен

Содержание лекционных занятий по дисциплине

Тема 1. Введение в технологии обработки и анализа данных

- 1.1 структура и обзор курса;
- 1.2 постановка задачи анализа данных, профессия Data Scientist;
- 1.3 обзор технологий анализа и обработки данных;
- 1.4 использование сводных таблиц в MS Excel, агрегатные функции.

Тема 2. Проектирование систем анализа и обработки данных

- 2.1 OLTP и OLAP: особенности, сравнительный анализ;
- 2.2 типовая архитектура систем обработки и анализа данных;
- 2.3 понятия Data Warehouse, Data Mart, ETL, Data Lake и др.;
- 2.4 обзор технологий и программного обеспечения.

Тема 3. Технология OLAP

- 3.1 назначение, классификация OLAP;
- 3.2 понятие многомерного OLAP куба;
- 3.3 подходы к организации структуры хранилища данных ROLAP: нормализация и денормализация таблиц, схема звезды, схема снежинки;
- 3.4 таблицы фактов и измерений, оси, меры: подходы и практики проектирования схемы OLAP-куба;
- 3.5 разработка схемы OLAP-куба Mondrian, язык описания схемы, знакомство с Pentaho Schema Workbench;
- 3.6 основы синтаксиса языка запросов MDX.

Тема 4. Понятие ETL. Разработка ETL-процесса с помощью инструмента Pentaho Data Integration

- 4.1 основные этапы и особенности реализации;
- 4.2 инструменты реализации ETL-процесса;
- 4.3 Pentaho Data Integration: библиотека компонентов для работы с данными, практики использования.

Тема 5. Язык статистического анализа данных R. Интегрированная среда разработки RStudio

- 5.1 назначение и возможности языка;
- 5.2 синтаксис, основные конструкции и библиотеки;
- 5.3 средства анализа временных рядов в R;
- 5.4 средства визуализации данных в R.

Тема 6. Методы машинного обучения: линейная регрессия. Прикладной анализ данных с помощью Python.

- 6.1 понятие линейной регрессии;
- 6.2 обзор синтаксиса языка Python и библиотек pandas, numpy, sklearn.

Тема 7. Методы машинного обучения: кластеризация. Прикладной анализ данных с помощью Python.

Тема 8. Методы машинного обучения: прогнозирование временных рядов. Прикладной анализ данных с помощью Python.

Содержание лабораторных занятий по дисциплине

Тема 1. Обзор и структурирование понятий и технологий в сфере анализа данных.

Предполагает построение ментальной карты (mind map) для понятия "Технологии анализа данных", а также знакомство с основными этапами анализа данных.

Тема 2. Быстрый анализ ограниченного массива данных с помощью инструмента "Сводная таблица (Pivot table)"

Предполагает знакомство на практике с основными этапами анализа данных на примере анализа предложенного датасета ограниченного размера с помощью сводных таблиц в MS Excel (или другом табличном процессоре). В процессе подготовки и анализа данных предусмотрено использование агрегатных и других функций Excel. Информация, полученная в результате анализа, должна быть представлена в виде оформленного аналитического отчета. Отчет должен содержать элементы визуализации данных и интерпретацию результатов анализа.

Тема 3. Технологии OLAP. Разработка многомерного куба на базе OLAP-сервера Mondrian.

Предполагает сравнительный анализ модели данных для OLTP и OLAP систем в рамках выбранной предметной области. В результате работы должна быть спроектирована и разработана схема OLAP-куба с помощью инструмента Pentaho Schema Workbench, входящего в состав комплекса средств Pentaho BI. Знакомство с синтаксисом языка запросов MDX.

Тема 4. Понятие ETL. Разработка ETL-процесса с помощью инструмента Pentaho Data Integration

Предполагает разработку механизма загрузки данных из CSV файла в реляционную БД (MySQL, PostgreSQL) с помощью библиотеки встроенных компонентов Pentaho Data Integration (Kettle). В качестве датасета и целевой БД должны использоваться исходные данные и хранилище, спроектированное в рамках лабораторной работы №3.

Тема 5. Прикладной анализ и визуализация данных с помощью языка R. Анализ временных рядов

Предполагает знакомство со средой разработки RStudio, статистический анализ датасета в формате CSV с визуализацией результатов средствами языка и платформы.

Тема 6. Прикладной анализ и визуализация данных с помощью языка Python. Линейная регрессия

Предполагает знакомство с синтаксисом языка Python, библиотеками pandas, numpy, sklearn для решения задач линейной регрессии

Тема 7. Прикладной анализ данных с помощью языка Python. Кластеризация

Предполагает знакомство с синтаксисом языка Python, библиотеками pandas, numpy, sklearn для решения задач кластеризации данных

Тема 8. Прикладной анализ данных с помощью языка Python. Прогнозирование временных рядов

Предполагает знакомство с синтаксисом языка Python, библиотеками pandas, numpy, sklearn для решения задач прогнозирования временных рядов.

5. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

В преподавании дисциплины «Технологии анализа данных» используются разнообразные образовательные технологии как традиционные, так и с применением активных и интерактивных методов обучения.

Активные и интерактивные методы обучения:

- интерактивные лекции с мультимедийным комплектом слайдов (темы № 1 – 8);
- разбор конкретных ситуаций (темы № 1 – 8);
- выполнение индивидуального лабораторного задания (темы № 1 – 8).

6. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Перечень контрольных вопросов и заданий для проведения текущего контроля

Рейтинг-контроль 1.

1. постановка задачи анализа данных, профессия Data Scientist;
2. обзор технологий анализа и обработки данных;
3. использование сводных таблиц в MS Excel, агрегатные функции;
4. OLTP и OLAP: особенности, сравнительный анализ;
5. типовая архитектура систем обработки и анализа данных;
6. понятия Data Warehouse, Data Mart, ETL, Data Lake и др.;
7. назначение, классификация OLAP;
8. понятие многомерного OLAP куба;
9. подходы к организации структуры хранилища данных ROLAP: нормализация и денормализация таблиц, схема звезды, схема снежинки;
10. таблицы фактов и измерений, оси, меры: подходы и практики проектирования схемы OLAP-куба;
11. основы синтаксиса языка запросов MDX.

Рейтинг-контроль 2.

1. понятие ETL. Разработка ETL-процесса с помощью инструмента Pentaho Data Integration;
2. основные этапы и особенности реализации;
3. инструменты реализации ETL-процесса;
4. Pentaho Data Integration: библиотека компонентов для работы с данными, практики использования;
5. язык статистического анализа данных R;
6. интегрированная среда разработки RStudio;
7. назначение и возможности языка R;
8. синтаксис, основные конструкции и библиотеки языка R;
9. средства анализа временных рядов в R;
10. средства визуализации данных в R.
11. методы машинного обучения: линейная регрессия;
12. прикладной анализ данных с помощью Python;
13. синтаксис языка Python и библиотек pandas, numpy, sklearn.

Рейтинг-контроль 3.

1. методы машинного обучения: кластеризация;
2. прикладной анализ данных с помощью Python;
3. библиотеки pandas, numpy, sklearn для решения задач кластеризации данных;
4. методы машинного обучения: прогнозирование временных рядов;
5. прикладной анализ данных с помощью Python;
6. библиотеки pandas, numpy, sklearn для решения задач прогнозирования временных рядов.

Промежуточная аттестация по итогам освоения дисциплины (экзамен)

1. постановка задачи анализа данных, профессия Data Scientist;
2. обзор технологий анализа и обработки данных;
3. использование сводных таблиц в MS Excel, агрегатные функции;

4. OLTP и OLAP: особенности, сравнительный анализ;
5. типовая архитектура систем обработки и анализа данных;
6. понятия Data Warehouse, Data Mart, ETL, Data Lake и др.;
7. назначение, классификация OLAP;
8. понятие многомерного OLAP куба;
9. подходы к организации структуры хранилища данных ROLAP: нормализация и денормализация таблиц, схема звезды, схема снежинки;
10. таблицы фактов и измерений, оси, меры: подходы и практики проектирования схемы OLAP-куба;
11. основы синтаксиса языка запросов MDX.
12. понятие ETL. Разработка ETL-процесса с помощью инструмента Pentaho Data Integration;
13. основные этапы и особенности реализации;
14. инструменты реализации ETL-процесса;
15. Pentaho Data Integration: библиотека компонентов для работы с данными, практики использования;
16. язык статистического анализа данных R;
17. интегрированная среда разработки RStudio;
18. назначение и возможности языка R;
19. синтаксис, основные конструкции и библиотеки языка R;
20. средства анализа временных рядов в R;
21. средства визуализации данных в R.
22. методы машинного обучения: линейная регрессия;
23. прикладной анализ данных с помощью Python;
24. синтаксис языка Python и библиотек pandas, numpy, sklearn.
25. методы машинного обучения: кластеризация;
26. прикладной анализ данных с помощью Python;
27. библиотеки pandas, numpy, sklearn для решения задач кластеризации данных;
28. методы машинного обучения: прогнозирование временных рядов;
29. прикладной анализ данных с помощью Python;
30. библиотеки pandas, numpy, sklearn для решения задач прогнозирования временных рядов.

Перечень заданий для самостоятельной работы студентов

1. OLTP и OLAP;
2. многомерный OLAP куб;
3. основы синтаксиса языка запросов MDX;
4. инструмент Pentaho Data Integration;
5. язык статистического анализа данных R;
6. методы машинного обучения;
7. язык Python;
8. библиотеки pandas, numpy, sklearn.

Самостоятельная работа обучающихся заключается в самостоятельном изучении отдельных тем, практической реализации типовых заданий по эти темам. Контроль выполнения самостоятельной работы проводится при текущих контрольных мероприятиях и на промежуточной аттестации по итогам освоения. Учебно-методическое обеспечение самостоятельной работы – основная литература [1-4].

Фонд оценочных средств для проведения аттестации уровня сформированности компетенций обучающихся по дисциплине оформляется отдельным документом.

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

7.1. Книгообеспеченность

№ п/п	Наименование литературы: автор, название, вид издания, издательство	Год издания	КНИГООБЕСПЕЧЕННОСТЬ	
			Количество экземпляров изданий в библиотеке ВлГУ в соответствии с ФГОС ВО	Наличие в электронной библиотеке ВлГУ
1	2	3	4	5
Основная литература				
1	Лесковец Ю., Анализ больших наборов данных / Лесковец Ю., Раджараман А., Джеффри Д. Ульман - М. : ДМК Пресс, 2016. - 498 с. - ISBN 978-5-97060-190-7	2016	-	http://www.studentlibrary.ru/book/ISBN9785970601907.html
2	Маккинли, У. Python и анализ данных / Уэс Маккинли ; пер. с англ. А.А. Слинкина. - Москва : ДМК Пресс, 2015. - 482 с. - ISBN 978-5-97060-315-4	2015	-	http://znanium.com/catalog/product/1027796
3	Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / С. Рашка ; пер. с англ. А.В. Логунова. - Москва : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0	2017	-	http://znanium.com/catalog/product/1027758
4	Шипунов А.Б., Наглядная статистика. Используем R! / А.Б. Шипунов, Е.М. Балдин, П.А. Волкова, А.И. Коробейников, С.А. Назарова, С.В. Петров, В.Г. Суфиянов. - М. : ДМК Пресс, 2012. - 298 с. - ISBN 978-5-94074-828-1	2012	-	http://www.studentlibrary.ru/book/ISBN9785940748281.html
Дополнительная литература				
1	Мхитарян, В.С. Анализ данных в MS Excel : учеб. пособие / В.С. Мхитарян, В.Ф. Шишов, А.Ю. Козлов. - М. : КУРС, 2019. - 368 с. - ISBN 978-5-906923-26-4	2019	-	http://znanium.com/catalog/product/1016934
2	Статистический анализ данных в MS Excel: Учебное пособие / Козлов А.Ю., Мхитарян В.С., Шишов В.Ф. - М.: НИЦ ИНФРА-М, 2016. - 320 с.: 60x90 1/16. - (Высшее образование: Бакалавриат) (Переплёт 7БЦ) ISBN 978-5-16-004579-5	2016	-	http://znanium.com/catalog/product/558444

7.2. Периодические издания:

1. Вестник компьютерных и информационных технологий ISSN 1810-7206

7.3. Интернет-ресурсы

1. www.edu.ru – портал российского образования
2. www.elbib.ru – портал российских электронных библиотек
3. www.eLibrary.ru – научная электронная библиотека
4. www.intuit.ru - интернет университета информационных технологий
5. library.vlsu.ru - научная библиотека ВлГУ

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Для реализации данной дисциплины имеются специальные помещения для проведения занятий лекционного типа, занятий лабораторного типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы. Лекции и лабораторные занятия проводятся в аудиториях, оборудованных мультимедийным проектором с экраном, с использованием комплекта слайдов, а также специализированным программным обеспечением (ауд. 424-2).


Перечень используемого лицензионного программного обеспечения:

- Операционная система Microsoft Windows 10
- Офисный пакет Microsoft Office 2016

Перечень условно бесплатного и свободно распространяемого ПО:

- Офисный пакет LibreOffice
- Платформа анализа данных Pentaho BI Server
- Среда разработки процессов интеграции данных Pentaho Data Integration
- Платформа и язык программирования Python, библиотеки sklearn, pandas, numpy
- Среда разработки программного обеспечения RStudio для языка программирования R
- Системы управления БД MySQL, PostgreSQL, Vertica (на выбор)
- Аналитическая платформа Tableau (опционально)

Рабочую программу составил: ст. преподаватель каф. ИСПИ Тимофеев А.А. 

Рецензент: к.т.н., генеральный директор ООО «Системный подход» Шориков А.В. 

Программа рассмотрена и одобрена на заседании кафедры ИСПИ

Протокол № 12 от 19.06.19 года.

Заведующий кафедрой  Жигалов И.Е.

Рабочая программа рассмотрена и одобрена на заседании учебно-методической комиссии
направления 09.04.04 «Программная инженерия»

Протокол № 12 от 19.06.19 года.

Председатель комиссии  Жигалов И.Е.

**ЛИСТ ПЕРЕУТВЕРЖДЕНИЯ
РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ**

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____

Рабочая программа одобрена на _____ учебный год.

Протокол заседания кафедры № _____ от _____ года.

Заведующий кафедрой _____