Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования

**Пермский национальный исследовательский
политехнический университет**

УТВЕРЖДАЮ
Проректор по учебной работе
_____ Н.В.Лобов
« 09 » декабря 20 19 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Дисциплина: Статистические методы анализа данных и технологии DataMining
(наименование)

Форма обучения: очная
(очная/очно-заочная/заочная)

Уровень высшего образования: магистратура
(бакалавриат/специалитет/магистратура)

Общая трудоёмкость: 180 (5)
(часы (ЗЕ))

Направление подготовки: 09.04.01 Информатика и вычислительная техника
(код и наименование направления)

Направленность: Высокопроизводительные вычислительные системы
(наименование образовательной программы)

Пермь 2019

1. Общие положения

1.1. Цели и задачи дисциплины

Цель дисциплины в формировании знаний, умений и навыков проведения самостоятельных исследований методами Data Mining и эффективного использования результатов уже готовых статистических исследований

Задачи дисциплины:

- знать методы и средства интеллектуального анализа данных;
- уметь самостоятельно использовать в практической деятельности интеллектуальный анализ данных с помощью информационных технологий
- уметь решать задачи статистического анализа с применением методов моделирования
- владеть навыками сбора и анализа информации в соответствующей профессиональной сфере, а также экспериментального исследования объектов профессиональной деятельности

1.2. Изучаемые объекты дисциплины

Математический инструментарий проведения сбора и анализа информации;
информационные технологии проведения интеллектуального анализа.

1.3. Входные требования

Не предусмотрены

2. Планируемые результаты обучения по дисциплине

Компетенция	Индекс индикатора	Планируемые результаты обучения по дисциплине (знать, уметь, владеть)	Индикатор достижения компетенции, с которым соотнесены планируемые результаты обучения	Средства оценки
ОПК-1	ИД-1ОПК-1	знает методы и средства интеллектуального анализа данных;	Знает основы высшей математики, физики, основы вычислительной техники и программирования	Защита лабораторной работы
ОПК-1	ИД-2ОПК-1	Умеет решать стандартные профессиональные задачи с применением методов математического моделирования	Умеет решать стандартные профессиональные задачи с применением естественнонаучных и общеинженерных знаний, методов математического анализа и моделирования	Защита лабораторной работы
ОПК-1	ИД-3ОПК-1	владеть навыками экспериментального исследования объектов профессиональной деятельности	Владеет навыками теоретического и экспериментального исследования объектов профессиональной деятельности	Защита лабораторной работы

Компетенция	Индекс индикатора	Планируемые результаты обучения по дисциплине (знать, уметь, владеть)	Индикатор достижения компетенции, с которым соотнесены планируемые результаты обучения	Средства оценки
ПКО-1	ИД-1ПКО-1	Знает порядок выявления результатов интеллектуальной деятельности	Знает порядок выявления охраноспособных объектов, определения соответствия выявленных результатов интеллектуальной деятельности условиям патентоспособности: задачи, подлежащие решению, технический результат, новизна объекта, изобрета-тельский уровень, промышленная применимость и прочее	Индивидуальное задание
ПКО-1	ИД-2ПКО-1	Владеет навыками сбора и анализа информации в соответствующей профессиональной сфере	Владеет навыками сбора и анализа информации об уровне научно-технического развития в соответствующей профессиональной сфере - поиска, отбора и анализа научно-технической, патентной, правовой информации	Индивидуальное задание
ПКО-1	ИД-2ПКО-1	умеет самостоятельно использовать в практической деятельности интеллектуальный анализ данных с помощью информационных технологий	Умеет самостоятельно приобретать и использовать в практической деятельности знания в области интеллектуальной собственности, в том числе с помощью информационных технологий	Защита лабораторной работы

3. Объем и виды учебной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		3
1. Проведение учебных занятий (включая проведение текущего контроля успеваемости) в форме: 1.1. Контактная аудиторная работа, из них:	62	62
- лекции (Л)	16	16
- лабораторные работы (ЛР)	16	16
- практические занятия, семинары и (или) другие виды занятий семинарского типа (ПЗ)	26	26
- контроль самостоятельной работы (КСР)	4	4
- контрольная работа		
1.2. Самостоятельная работа студентов (СРС)	82	82
2. Промежуточная аттестация		
Экзамен	36	36
Дифференцированный зачет		
Зачет		
Курсовой проект (КП)		
Курсовая работа (КР)		
Общая трудоемкость дисциплины	180	180

4. Содержание дисциплины

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
3-й семестр				

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
Подготовка и предварительный анализ данных, введение в Data Mining,	4	4	6	20
Общая концепция методологии Data Mining и технологии реализации. Обзор задач, решаемых методами Data Mining. Классификация методов Data Mining по различным признакам. Этапы интеллектуального анализа данных: анализ предметной области, постановка задачи, подготовка данных. Процесс подготовки данных, понятия качества данных, грязных данных, этапы очистки данных.Этапы процесса Data Mining, связанные с построением, проверкой, оценкой, выбором и коррекцией моделей. Процесс Data Mining как последовательность этапов и как последовательность работ, выполняемых исполнителями ролей Data Mining. Подготовка и предварительный анализ данных Анализ взаимосвязей (корреляций) между переменными – коэффициенты корреляции Пирсона, Спирмена и Кендалла. Сравнение групп – критерии Стьюдента, Манна – Уитни и дисперсионный анализ (ANOVA). Инструменты очистки и редактирования данных, основные функции инструментов очистки данных, классификация ошибок в данных, которые возникают в результате использования средств очистки данных. Инструменты очистки данных.				
Типовые задачи Data Mining и методы их решения	6	6	10	30
Задачи классификации и методы решения. Постановка задач, ключевые понятия и определения. Методы, применяемые для решения задач классификации: индукция деревьев решений; байесовские сети (Bayesian Networks); нейронные сети (neural networks);статистические методы, в частности, линейная регрессия.Преимущества деревьев решений. Интуитивность деревьев решений. Точность. Процесс конструирования дерева решений. Метод "ближайшего соседа". Регрессионный анализ. Последовательность этапов регрессионного анализа. Классические методы регрессионного анализа: множественная и логистическая регрессии, выбор переменных для анализа. Установление формы зависимости. Определение функции регрессии. Оценка неизвестных значений зависимой переменной. Уравнение регрессии. Случайное отклонение. Коэффициент множественной корреляции R Сравнение качества построенных моделей. Задачи кластеризации: постановка задачи, ключевые понятия и определения, метод k-средних				

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
и EM-алгоритм. Задача понижения размерности, метод независимых компонент (independent component analysis). Меры сходства. Иерархический кластерный анализ в SPSS. Алгоритм k-средних (k-means). Классические методы кластеризации – Метод иерархической кластеризации (tree clustering). Классические методы понижения размерности: метод главных компонент (principal component analysis), факторный анализ (Factor Analysis). Прогнозирование временных рядов – основные понятия (тренд, сезонность, календарные эффекты, разложение ряда), классическая модель АРПСС (ARIMA), экспоненциальное сглаживание, анализ лагов. Нейронные сети (Neural networks): элементы нейронных сетей, обучение нейронных сетей, модели нейронных сетей, программное обеспечение для работы с нейронными сетями. Пакет Matlab.				
Методы анализа данных и используемые приложения	6	6	10	32
Задачи поиска ассоциативных правил. Методы поиска ассоциативных правил. Алгоритм Apriori. Приложения с применением ассоциативных правил. Задачи и методы визуализации. Способы представления информации в одно-, двух-, трехмерном измерениях и более. Принципы качественной визуализации. Основные тенденции в области визуализации. Визуализация инструментов Data Mining. Визуализация Data Mining моделей. Приложения Data Mining и используемые программные продукты СППР, их типы и компоненты. OLAP-технологии, архитектуры OLAP-серверов, интеграции Data Mining и OLAP. Технология хранилищ данных и преимущества их использования для процесса Data Mining. Процесс Data Mining и методологии CRISP и SEMMA. Программное обеспечение Data Mining для решения задач оценивания и прогнозирования. Система STATISTICA Data Miner, средства анализа и схема работы.				
ИТОГО по 3-му семестру	16	16	26	82
ИТОГО по дисциплине	16	16	26	82

Тематика примерных практических занятий

№ п.п.	Наименование темы практического (семинарского) занятия
1	Этапы интеллектуального анализа данных: анализ предметной области, постановка задачи, подготовка данных, понятия качества данных, грязных данных, этапы очистки данных .
2	Процесс Data Mining как последовательность этапов и как последовательность работ, выполняемых исполнителями ролей Data Mining.
3	Анализ взаимосвязей (корреляций) между переменными – коэффициенты корреляции Пирсона, Спирмена и Кендалла.
4	Модели дисперсионного анализа (ANOVA).
5	Выбор вида зависимости, набора значимых факторов, проверка на значимость.
6	Методы, применяемые для решения задач классификации: индукция деревьев решений; байесовские сети (Bayesian Networks); нейронные сети (neural networks)
7	Методы, применяемые для решения задач кластеризации: метод k-средних и EM-алгоритм. Задача понижения размерности. Метод независимых компонент (independent component analysis).
8	Система STATISTICA Data Miner, средства анализа и схема работы.

Тематика примерных лабораторных работ

№ п.п.	Наименование темы лабораторной работы
1	Анализ данных, представленных в виде временных рядов.
2	Анализ взаимосвязей между переменными (корреляционный анализ).
3	Регрессионный анализ: установление вида связи и проверка на качество.
4	Иерархический кластерный анализ в SPSS. Итеративная кластеризация в SPSS.

5. Организационно-педагогические условия

5.1. Образовательные технологии, используемые для формирования компетенций

Проведение лекционных занятий по дисциплине основывается на активном методе обучения, при которой учащиеся не пассивные слушатели, а активные участники занятия, отвечающие на вопросы преподавателя. Вопросы преподавателя нацелены на активизацию процессов усвоения материала, а также на развитие логического мышления. Преподаватель заранее намечает список вопросов, стимулирующих ассоциативное мышление и установления связей с ранее освоенным материалом.

Практические занятия проводятся на основе реализации метода обучения действием: определяются проблемные области, формируются группы. При проведении практических занятий преследуются следующие цели: применение знаний отдельных дисциплин и креативных методов для решения проблем и приятия решений; отработка у обучающихся навыков командной работы, межличностных коммуникаций и развитие лидерских качеств; закрепление основ теоретических знаний.

Проведение лабораторных занятий основывается на интерактивном методе обучения, при котором обучающиеся взаимодействуют не только с преподавателем, но и друг с другом. При этом доминирует активность учащихся в процессе обучения. Место преподавателя в интерактивных занятиях сводится к направлению деятельности обучающихся на достижение целей занятия.

При проведении учебных занятий используются интерактивные лекции, групповые дискуссии, ролевые игры, тренинги и анализ ситуаций и имитационных моделей.

5.2. Методические указания для обучающихся по изучению дисциплины

При изучении дисциплины обучающимся целесообразно выполнять следующие рекомендации:

1. Изучение учебной дисциплины должно вестись систематически.
2. После изучения какого-либо раздела по учебнику или конспектным материалам рекомендуется по памяти воспроизвести основные термины, определения, понятия раздела.
3. Особое внимание следует уделить выполнению отчетов по практическим занятиям, лабораторным работам и индивидуальным комплексным заданиям на самостоятельную работу.
4. Вся тематика вопросов, изучаемых самостоятельно, задается на лекциях преподавателем. Им же даются источники (в первую очередь вновь изданные в периодической научной литературе) для более детального понимания вопросов, озвученных на лекции.

6. Перечень учебно-методического и информационного обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Печатная учебно-методическая литература

№ п/п	Библиографическое описание (автор, заглавие, вид издания, место, издательство, год издания, количество страниц)	Количество экземпляров в библиотеке
1. Основная литература		
1	Берикашвили В. Ш. Статистическая обработка данных, планирование эксперимента и стучайные процессы : учебное пособие для бакалавриата и магистратуры / В. Ш. Берикашвили, С. П. Оськин. - Москва: Юрайт, 2019.	6
2	Методы и модели анализа данных: OLAP и Data Mining : учебное пособие / А. А. Барсегян [и др.]. - Санкт-Петербург: БХВ-Петербург, 2004.	12

2. Дополнительная литература		
2.1. Учебные и научные издания		
1	Репин С. В. Математические методы обработки статистической информации с помощью ЭВМ : пособие для исследователей гуманитарных специальностей / С. В. Репин, С. А. Шеин. - Минск: Универститетское, 1990.	6
2.2. Периодические издания		
	Не используется	
2.3. Нормативно-технические издания		
	Не используется	
3. Методические указания для студентов по освоению дисциплины		
1	Бродягин В. В. Основы компьютерных технологий решения геологических задач : учебное пособие / В. В. Бродягин. - Пермь: Изд-во ПГТУ, 2008.	29
4. Учебно-методическое обеспечение самостоятельной работы студента		
1	Дубнов П. Ю. Обработка статистической информации с помощью SPSS / П. Ю. Дубнов. - Москва: АСТ, NT Press, 2004.	4

6.2. Электронная учебно-методическая литература

Вид литературы	Наименование разработки	Ссылка на информационный ресурс	Доступность (сеть Интернет / локальная сеть; авторизованный / свободный доступ)
Методические указания для студентов по освоению дисциплины	Бродягин В. В. Основы компьютерных технологий решения геологических задач : учебное пособие / В. В. Бродягин. - Пермь: Изд-во ПГТУ, 2008.	http://elib.pstu.ru/Record/RUPSTUbooks130806	сеть Интернет; авторизованный доступ

6.3. Лицензионное и свободно распространяемое программное обеспечение, используемое при осуществлении образовательного процесса по дисциплине

Вид ПО	Наименование ПО
Операционные системы	MS Windows 8.1 (подп. Azure Dev Tools for Teaching)
Операционные системы	Windows 10 (подп. Azure Dev Tools for Teaching)
Офисные приложения.	Microsoft Office Professional 2007. лиц. 42661567
Прикладное программное обеспечение общего назначения	MATLAB 7.9 + Simulink 7.4 Academic, ПНИПУ 2009 г.
Системы управления проектами, исследованиями, разработкой, проектированием, моделированием и внедрением	IBM SPSS Statistic Base
Среды разработки, тестирования и отладки	Microsoft Visual Studio (подп. Azure Dev Tools for Teaching)

Вид ПО	Наименование ПО
Среды разработки, тестирования и отладки	Среда разработки RStudio
Среды разработки, тестирования и отладки	Язык R

6.4. Современные профессиональные базы данных и информационные справочные системы, используемые при осуществлении образовательного процесса по дисциплине

Наименование	Ссылка на информационный ресурс
База данных Scopus	https://www.scopus.com/
База данных Web of Science	http://www.webofscience.com/
База данных научной электронной библиотеки (eLIBRARY.RU)	https://elibrary.ru/
Научная библиотека Пермского национального исследовательского политехнического университета	http://lib.pstu.ru/
Электронно-библиотечеая система Лань	https://e.lanbook.com/
Электронно-библиотечная система IPRbooks	http://www.iprbookshop.ru/
Электронно-библиотечная система IPRbooks	http://www.iprbookshop.ru/
Электронно-библиотечная система IPRbooks	http://www.iprbookshop.ru/
Виртуальный читальный зал Российской государственной библиотеки	https;//dvs.rsl.ru/
Виртуальный читальный зал Российской государственной библиотеки	https;//dvs.rsl.ru/
Электронная библиотека диссертаций Российской государственной бибилиотеки	http://www.diss.rsl.ru/

7. Материально-техническое обеспечение образовательного процесса по дисциплине

Вид занятий	Наименование необходимого основного оборудования и технических средств обучения	Количество единиц
Лабораторная работа	Компьютерный класс	10
Лекция	Лекционная аудитория: проектор и компьютер	1
Практическое занятие	Компьютерный класс	10

8. Фонд оценочных средств дисциплины

Описан в отдельном документе