

## АННОТАЦИЯ ДИСЦИПЛИНЫ

«Статистические методы анализа данных и технологии DataMining»

Дисциплина «Статистические методы анализа данных и технологии DataMining» является частью программы магистратуры «Разработка программно-информационных систем» по направлению «09.04.04 Программная инженерия».

### **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.

### **Изучаемые объекты дисциплины**

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

### 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:	62	62	
1.1. Контактная аудиторная работа, из них:			
- 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	

### Краткое 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
Типовые задачи Data Mining и методы их решения	6	6	10	30
<p>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".</p> <p>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.</p> <p>Задачи кластеризации: постановка задачи, ключевые понятия и определения, метод k-средних и EM-алгоритм. Задача понижения размерности, метод независимых компонент (independent component analysis). Меры сходства. Иерархический кластерный анализ в SPSS. Алгоритм k-средних (k-means).</p> <p>Классические методы кластеризации – Метод иерархической кластеризации (tree clustering). Классические методы понижения размерности: метод главных компонент (principal component analysis), факторный анализ (Factor Analysis).</p> <p>Прогнозирование временных рядов – основные понятия (тренд, сезонность, календарные эффекты, разложение ряда), классическая модель АРПСС (ARIMA), экспоненциальное сглаживание, анализ лагов.</p> <p>Нейронные сети (Neural networks): элементы нейронных сетей, обучение нейронных сетей, модели нейронных сетей, программное обеспечение для работы с нейронными сетями. Пакет Matlab.</p>				

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.				
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,				

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