Программа учебной дисциплины «Анализ и визуализация данных»

Утверждена
Академическим руководителем
«26» июня 2018 г.

Автор	Милков М.А.
Число кредитов	3
Контактная работа (час.)	48
Самостоятельная работа (час.)	66
Курс	1
Формат изучения дисциплины	full time

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Анализ и визуализация данных» являются овладение студентами основных методов и инструментов для проведения анализа и визуализации данных.

В результате освоения дисциплины студент должен:

знать:

- основные принципы и этапы проведения анализа данных;
- типы задач, решаемые с помощью анализа данных;
- основные методы анализа данных, границы их применимости.

уметь:

- определять подходящий(ие) метод(ы) для решения конкретной задачи;
- оценивать степень достоверности результатов, полученных с помощью статистических методов исследования;
- толковать смысл полученных результатов.

владеть:

- навыками использования инструментов для обработки, анализа и визуализации данных;
- навыками применения основных методов статистического анализа для решения производственных задач;
- навыками применения основных методов машинного обучения для решения производственных задач.

Изучение дисциплины «Физика» базируется на следующих дисциплинах:

- математическая статистика;
- теория вероятности.

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- знать основные термины статистики и теории вероятности;
- обладать навыками работы с табличным программным обеспечением.

Основные положения дисциплины могут быть использованы в дальнейшем при изучении следующих дисциплин: научно-исследовательский семинар и выполнение выпускной квалификационной работы.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в анализ данных и Statistica.

Что такое наука о данных. Обзор реальных задач, в которых применимы методы анализа данных.

Интерфейс Statistica. Основные типы объектов Statistica. Устройство таблиц Statistica. Типы данных и типы шкал. Использование формул.

Импорт данных. Табличные файлы. Построение запросов к реляционным базам данных. Процедуры ETL: извлечение, преобразование, загрузка.

Предобработка данных. Поиск дублей. Поиск разреженных данных. Поиск и обработка выбросов.

Описательные статистики. Точечное и интервальное оценивание. Параметры положения. Параметры разброса. Квантили.

Визуализация. Типы графиков. Одно-, двух-, трех-мерные графики. Категоризованные графики. Настройка отображения графиков.

Тема 2. Поиск связей.

Кореляции. Корреляция Пирсона. Понятие статистической значимости. Параметрические и непараметрические методы. Корреляция Спирмена. Частные и ложные корреляции.

Сравнение средних в двух группах. Т-критерий. Независимые и зависимые выборки. Ограничения применимости. Непараметрические аналоги.

Сравнение средних в нескольких группах. Дисперсионный анализ. Главные эффекты и эффекты взаимодействия. Контрасты.

Связь качественных признаков. Таблицы сопряженности. Критерий Хи-квадрат.

Регрессии. Парная регрессия. Множественная линейная регрессия. Метод наименьших квадратов. Пошаговые процедуры. Проблема мультиколлениарности. Нелинейное оценивание.

Классификация. Бинарная классификация. Метод Логит-регрессии.

Тема 3. Многомерный анализ.

Кластеризация. Методы к-средних, иерархический, DBSCAN.

Снижение размерности. Факторы. Метод главных компонент. Вращение признаков.

Множественная классификация. Дискриминантный анализ Фишера. Пошаговые процедуры.

Тема 4. Анализ временных рядов.

Временные ряды и кроссекционные данные. Автокорреляционная функция. Сезонность. Тренды.

Прогнозирование временных рядов. Методы сезонной декомпозиции, экспоненциальное сглаживание, АРПСС.

Тема 5. Методы машинного обучения.

Недостатки классических методов анализа данных. Концепция Data Mining. Методы машинного обучения.

Искусственные нейронные сети. Биологический принцип и математическая интерпретация. Многослойный персептрон. Обучение нейронных сетей. Проблема переобучения.

Решающие деревья. Алгоритмы ветвления. Правила остановки обучения. Обрезание дерева. Дерево CART. Случайные леса. Градиентный бустинг деревьев.

III. ОЦЕНИВАНИЕ

Оценки по всем формам текущего и итогового контроля выставляются по 10-ти балльной шкале.

Преподаватель оценивает работу студентов на семинарских и практических занятиях:

оценивается активность студента в дискуссиях, правильность решения задач, уровень ориентированности студента в демонстрируемых им программах, понимание сильных сторон и ограничений используемых инструментов. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость. Оценка по 10-ти балльной шкале за работу на практических занятиях определяется перед итоговым контролем — Оауд.

Преподаватель оценивает самостоятельную работу студентов (задания, которые выдаются на семинарских занятиях). Оценки за самостоятельную работу студента преподаватель выставляет в рабочую ведомость. Оценка по 10-ти балльной шкале за работу на практических занятиях определяется перед итоговым контролем — Осам.

Накопленная оценка по дисциплине рассчитывается по формуле:

$$Онакопл = 0.5*Оауд+ 0.5*Осам$$

Способ округления накопленной оценки: в пользу студента.

В диплом выставляется результирующая оценка по учебной дисциплине, которая вычисляется по следующей формуле:

$$Орезультат = 0.5*Онакопл+ 0.5*Оэкзамен$$

Способ округления результирующей оценки: в пользу студента.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Примеры вопросов для самопроверки студентов:

- Отличие параметрических и непараметрических методов анализа данных?
- Критерии применимости Т-критерия.
- Как автоматизировано отобрать переменные для модели множественной регрессии?
- В чем отличие и что общего у задач регрессии и классификации?
- Какой метод машинного обучения предпочтителен для решения задачи классификации?

Вопросы для оценки качества освоения дисциплины
Примерный перечень вопросов к экзамену по всему курсу:

- Что такое частная корреляция?
- Что такое проблема мультиколлениарности? Перечислите способы решения данной проблемы.
- Как оценивается статистическая значимость модели множественной регрессии?
- Что такое коэффициент детерминации?
- В чем отличие классификации и кластерного анализа?
- Методы кластерного анализа.
- В чем различие и что общего у методов дискриминантного анализа и деревьев решений?
- Что такое проблема переобучения? Какие способы решения данной проблемы?

V. РЕСУРСЫ

5.1 Основная литература

1. Орлов, А.И. Вероятность и прикладная статистика: основные факты: справ. / А. И. Орлов. – М.: КноРус, 2010. – 190 с.
2. Тернер, Д. Вероятность, статистика и исследование операций / Д. Тернер; Пер. с англ. Е. З. Демиденко, В. С. Занадворова; Под ред. А. А. Рывкина. – М.: Статистика, 1976. – 432 с.
3. Боровиков, В. П. Популярное введение в современный анализ данных в системе Statistica: методология и технология современного анализа данных : уч. пособие / В. П. Боровиков. – М.: Горячая линия-Телеком, 2013. – 288 с.

5.2 Дополнительная литература

1. http://statisitca.ru Портал примеров анализа и визуализации данных с использование ПО Statistica.
2. Боровиков В.П. Популярное введение в современный анализ данных в системе STATISTICA, Горячая линия-Телеком, 2013.
3. Ивченко Г.И., Медведев Ю.И. Математическая статистика, Высшая школа, 1992.
4. Розанов Ю.А. Теория вероятностей, случайные процессы и математическая статистика, Наука 1985.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*

5.4 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:
- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.