Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и
образовательным инновациям
_____ О. Г. Прохоренко
30 июня 2023 г.

Регистрационный № УД –12644/уч.

ТЕХНОЛОГИИ АНАЛИЗА И ВИЗУАЛИЗАЦИИ ДАННЫХ

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:

1-31 03 04Информатика

2023г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО1-31 0304-2021, типового учебного плана №G 31-1-029/пр.-тип. от 30.06.2021 и учебного плана БГУ №G 31-1-031/уч. от 30.06.2021.

СОСТАВИТЕЛИ:

М.М. Лукашевич –доцент кафедры информационных систем управления Белорусского государственного университета, кандидат технических наук, доцент

РЕЦЕНЗЕНТ:

А.М. Белоцерковский – заведующий отделом интеллектуальных информационных систем ГНУ «Объединенный институт проблем информатики» Национальной академии наук Беларуси», кандидат технических наук, доцент

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой информационных систем управления
Белорусского государственного университета
(протокол № 18 от 08.06.2023 г.).

Научно-методическим Советом БГУ
(протокол № 9 от 29.06.2023 г.)

Заведующий кафедрой _____ В.В. Краснопрошин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная дисциплина «Технологии анализа и визуализации данных» ориентирована на обучение студентов алгоритмическому обеспечению систем анализа и интерпретации данных, знакомит студентов с основами визуального восприятия человека и типами визуализации в зависимости от используемых данных. Существенное внимание уделено вопросам классификации и кластеризации данных с использованием детерминированных и статистических моделей. Рассмотрены методы снижения размерностей данных. Изучаются интегрированные среды разработки и современные пакеты программ для анализа данных.

Основой для обучения являются компетенции, сформированные при изучении дисциплины «Теория вероятности и математическая статистика».

Цели и задачи учебной дисциплины

Цель преподавания учебной дисциплины «Технологии анализа и визуализации данных» –сформировать у студента теоретические и практические компетенции в области целостного представления, понимания места и роли, применения технологий анализа данных, ознакомить студентов с основным принципам визуализации разных типов данных, дать практические навыки визуализации.

При изложении курса важно показать возможности использования интегрированных сред разработки и пакетов прикладных программ для решения прикладных задач, возникающих в различных областях науки, техники, экономики и производства.

Задачи учебной дисциплины

Основной **задачей**, решаемой при изучении учебной дисциплины «Технологии анализа и визуализации данных», является подготовка специалиста, владеющего знаниями в области прикладных технологий анализа и визуализации данных; понимающего концепции и технологии современного анализа и визуализации данных; владеющего умениями и навыками самостоятельного решения задач анализа и визуализации данных с использованием современных программных средств,

Место учебной дисциплины в системе подготовки специалиста с высшим образованием

Учебная дисциплина относится к **циклу** дисциплин специализаций компонента учреждения высшего образования учебного плана специальности **1-31 03 04 Информатика.**

Программа составлена с учётом межпредметных **связей** с учебными дисциплинами.

Дисциплина «Технологии анализа данных» непосредственно связана с дисциплинами:

– «Дискретная математика и математическая логика», «Вычислительные методы алгебры», «Методы оптимизации». Теоретические основы, излагаемые в указанных дисциплинах, используются при реализации и оценке эффективности алгоритмов анализа данных, исследовательском анализе данных.

Сформированные при изучении дисциплины «Технологии анализа и визуализации данных» компетенции являются основой для изучения дисциплины «Интеллектуальные информационные системы».

Знания, полученные в учебной дисциплине, используются при изучении дисциплин специализации, при выполнении курсовых и дипломных работ, а также используются как инструментарий для моделирования и компьютерного решения задач ряда математических дисциплин, изучаемых на старших курсах.

Требования к компетенциям

Освоение учебной дисциплины «Технологии анализа и визуализации данных» должно обеспечить формирование следующей **универсальной компетенции:**

УК-2. Решать стандартные задачи профессиональной деятельности на основе применения информационно-коммуникационных технологий.

В результате освоения учебной дисциплины студент должен:

знать:

─методы и технологии разведочного анализа данных;

─методы и технологии кластерного анализа и классификации данных;

─базовые программно-технические решения анализа данных;

─методы и технологии визуализации данных;

─базовые программно-технические решения визуализации данных;

уметь:

─ программировать задачи анализа данных средней сложности;

─ оценивать эффективность различных алгоритмических решений анализа данных;

─ разрабатывать программные приложения анализа данных с заданной функциональностью;

─ реализовывать задачи визуализации данных средней сложности;

─ оценивать эффективность различных решений визуализации данных;

─ разрабатывать программные приложения визуализации данных с заданной функциональностью;

владеть:

– навыками разработки программных модулей на основе прикладных пакетов для анализа данных;

– навыками программирования анализа данных,

– навыками работы с программными модулями и средствами для визуализации данных;

– навыками визуализации данных.

Структура учебной дисциплины

Форма получения высшего образования – дневная (очная).

Дисциплина изучается в 5 семестре. Всего на изучение учебной дисциплины «Технологии анализа и визуализации данных» отведено:

– для очной формы получения высшего образования – 108 часов, в том числе 68 аудиторных часа, из них: лекции – 34 часа, лабораторные работы – 30 часов, управляемая самостоятельная работа – 4 часа.

Трудоемкость учебной дисциплины составляет 3 зачетных единицы.

Форма промежуточной аттестации – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Введение в технологии анализа и визуализации данных

Тема 1.1. *Основные понятия и определения*

От баз данных к анализу данных. Определение анализа данных. Основные понятия. Типы задач: обучение с учителем (регрессия, классификация), обучение без учителя (кластеризация, поиск аномалий, снижение размерностей), частичное обучение, обучение с подкреплением. История визуализации данных. Основные концепции визуального восприятия графиков. Разбор типов данных и выбор подходящих под них графиков. Работа с количественными и качественными данными.

Тема 1.2. *Интегрированные среды разработки и пакеты программ*

Библиотеки научных вычислений: NumPy, SciPy, Scikit-learn. Библиотека для работы с табличными данными Pandas. Среда интерактивных вычислений JupyterNotebook: настройка и установка, основные принципы работы.

Онлайн и оффлайн сервисы визуализации данных. Библиотеки визуализации: Matplotlib, Seaborn. Применение pandasв задаче исследования данных. Виды графиков и диаграмм. Основные элементы диаграммы. Создание диаграммы. Форматы изображений. График функции. Гистограмма.

Раздел 2. Прикладные технологии анализа и визуализации данных

Тема 2.1.*Задачи классификации и регрессии*

Математическая постановка задач регрессии и классификации. Метрические методы регрессии и классификации: метод ближайших соседей, взвешенный метод ближайших соседей. Метод опорных векторов в задаче классификации. Ядерный переход. Метод опорных векторов для задачи регрессии. Деревья решений для задачи классификации. Алгоритмы построения деревьев. Ансамбли моделей. Случайный лес.

Тема 2.2. *Кластерный анализ*

Математическая модель кластеризации. Меры однородности объектов. Расстояние между объектами. Меры близости между кластерами. Методы кластерного анализа. Иерархические агломеративные методы. Последовательные кластер-процедуры. Метод K-средних. Графическое представление результатов кластер-анализа.

Тема 2.3. *Работа с признаками*

Предобработка признаков. Категориальные признаки. Разреженные признаки. Отбор признаков

Тема 2.4. *Метрики качества моделей. Выбор модели*

Метрики в задаче регрессии: MAE, MSE, MAPE, R 2. Метрики в задаче классификации: кросс-энтропия, precision, recall, F-мера, ROC-кривая, AUC ROC. Обобщающая способность модели и ее оценка: отложенная выборка, кросс-валидация. Визуализация моделей классификации, регрессии, кластерного анализа, результатов понижения размерности. Визуальное представление оценок качества моделей классификации, регрессии, кластерного анализа, результатов понижения размерности.

Тема 2.5. *Понижение размерности данных*

Сущность главных компонент, их свойства, геометрическая интерпретация. Графическое представление результатов. Аппроксимация с использованием главных компонент.

Тема 2.6.*Прикладные задачи анализа и визуализации данных*

Основные области: компьютерное зрение, обработка естественного языка, рекомендательные системы, анализ временных рядов, обучение ранжированию, построение выводов по данным. Визуализация геопространственных данных. Практические примеры.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением электронных средств обучения (ДОТ)

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов				Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия		
1	2	3	4	5	6	8	9
	Технологии анализа и визуализации данных	**34**			**30**	**4**	
1.	**Введение в технологии анализа и визуализации данных**	**4**			**4**		
1.1.	Основные понятия и определения	2			2		Экспресс-опрос
1.2.	Интегрированные среды разработки и пакеты программ	2			2		Экспресс-опрос
2.	**Прикладные технологии анализа и визуализации данных**	**30**			**26**	**4**	
2.1.	Задачи классификации и регрессии	8			6		Коллоквиум по материалам тем 1.1, 1.2, 2.1
2.2.	Кластерный анализ	6			4		Расчетно-графическая работа №1
2.3.	Работа с признаками	4			4		Расчетно-графическая работа №2
2.4.	Метрики качества моделей. Выбор модели	4			4		Расчетно-графическая работа №3
2.5.	Понижение размерности данных	4			4	2	Экспресс-опрос
2.6.	Прикладные задачи анализа и визуализации данных	4			4	2	Расчетно-графическая работа №4

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Маккинни, У. Python и анализ данных / У. Маккинни; перевод с английского А. А. Слинкина. – 2-ое изд., испр. и доп. – Москва: ДМК Пресс, 2020. – 540 с.
2. Грас Д., DataScience. Наука о данных с нуля: Пер. с англ. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2021. – 416 с.: ил.
3. Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn, Keras и TensorFlow / О. Жерон. – 2-ое изд. - Диалектика, 2020. – 1040 с.
4. Горелик, А. Корпоративное озеро больших данных: новый подход к использованию BigData и DataSciencев бизнесе / А. Горелик. – Москва: Эксмо, 2023. – 272 с.
5. Манцер, Т. Визуализация данных: полный курс для начинающих специалистов / Т. Манцер. – Москва: Эксмо, 2023. – 464 с.

Перечень дополнительной литературы

1. Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения: руководство / С. Рашка; перевод с английского А. В. Логунова. – Москва: ДМК Пресс, 2017. – 418 с.
2. Говори на языке диаграмм: пособие по визуальным коммуникациям/ДжинЖелязны; пер.с англ. А.Мучника и Ю.Корнилович – 6-е изд. – М.: Манн, Иванов и Фербер, 2016. – 304 с.
3. Барсегян, А.А., Куприянов М. С. и др. Технологии анализа данных: DataMining, VisualMining, TextMinning, OLAP: учеб. – 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2007.
4. Барсегян, А.А., Куприянов М. С. и др. Методы и модели анализа данных: OLAP и DataMining: учеб. пособие – СПб.: БХВ-Петербург, 2004.

Электронные ресурсы

1. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru.
2. Онлайн средство визуализации данных. [Электронный ресурс]. – Режим доступа: https://www.gapminder.org/

Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. *Устная форма:* экспресс-опрос.
2. *Письменная форма:* коллоквиум.
3. *Устно-письменная форма:* расчетно-графические работы.

В качестве рекомендуемых технических средств диагностики используетсяобучение, организованное на платформе Moodle (https://edufpmi.bsu.by).

Формой промежуточной аттестации по дисциплине учебным планом предусмотрен экзамен.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку (формирование оценки за текущую успеваемость):

- отчёты по расчетно-графическим работам – 35 %;
- коллоквиум – 30 %;
- устный опрос – 35%.

Итоговая отметка по дисциплине рассчитывается на основе отметки текущей успеваемости и экзаменационной отметки с учетом их весовых коэффициентов. Вес отметки по текущей успеваемости составляет 40 %, экзаменационной – 60 %.

Рекомендуемая тематика коллоквиума

1) Коллоквиум «Специализированные пакеты numpy и pandas».

Примерный перечень заданий для управляемой самостоятельной работы студентов

Тема 2.5. Понижение размерности данных (2 ч.)

1. Метод главных компонент (Principal Component Analysis, PCA). Форма контроля – экспресс-опрос.
2. Метод t-SNE (t-distributed Stochastic Neighbor Embedding). Форма контроля – экспресс-опрос.

Тема 2.6. Прикладные задачи анализа и визуализации данных (2 ч.)

1. Основы визуализации данных и визуального восприятия. Форма контроля – экспресс-опрос.

2. Выбор алгоритмов визуализации данных под разные типы данных. Форма контроля – экспресс-опрос.

Примерная тематика лабораторных занятий

Занятие № 1. Настройка среды разработки и окружения, основы работы с GPU.

Занятие № 2. Разведочный анализа данных и их визуализация.

Занятие № 3. Решение задачи классификации данных.

Занятие № 4. Решение задачи регрессии.

Занятие № 5. Использование визуализации при решении задач классификации и регрессии.

Занятие № 6-7. Решение задачи кластеризации данных. Использование визуализации при кластеризации данных. Расчетно-графическая работа №1.

Занятие № 8. Предобработка и анализ информативности признаков.

Занятие № 9-10. Отбор и разработка новых признаков. Оценка качества классификации данных. Расчетно-графическая работа №2.

Занятие № 11. Оценка качества регрессии и кластеризации данных.

Занятие № 12-13. Понижение размерности данных. Визуализация результатов понижения размерности. Расчетно-графическая работа №3.

Занятие № 14-15. Решение прикладных задач анализа данных. Визуализация геопространственных данных. Расчетно-графическая работа №4.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используются следующие методы:

– ***метод группового обучения,*** который представляет собой форму организации учебно-познавательной деятельности обучающихся, предполагающую функционирование разных типов малых групп, работающих как над общими, так и специфическими учебными заданиями.

В качестве технических средств для организации работы в рамках учебной дисциплины рекомендуется использовать Образовательный портал БГУ (https://edufpmi.bsu.by) – инструмент с эффективной функциональностью контроля, тренинга и самостоятельной работы.

– ***практико-ориентированный подход,*** который предполагает освоение содержания образования через решения практических задач; приобретение навыков эффективного выполнения разных видов профессиональной деятельности; ориентацию на генерирование идей, реализацию групповых студенческих проектов; использование процедур, способов оценивания, фиксирующих профессиональные компетенции.

Методические рекомендации по организации самостоятельной работы обучающихся

Для организации самостоятельной работы студентов по учебной дисциплине следует использовать современные информационные ресурсы: разместить на образовательном портале комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, методические указания к лабораторным занятиям, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к зачету, задания, тесты, вопросы для самоконтроля и др., список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к экзамену

1. Определение анализа данных.
2. Типы задач: обучение с учителем (регрессия, классификация).
3. Типы задач: обучение без учителя (кластеризация, поиск аномалий, снижение размерностей.
4. Среда интерактивных вычислений JupyterNotebook: настройка и установка, основные принципы работы.
5. Метод ближайших соседей, взвешенный метод ближайших соседей.
6. Метод опорных векторов в задачах классификации и регрессии.
7. Деревья решений для задачи классификации. Алгоритмы построения деревьев.
8. Ансамбли моделей. Случайный лес.
9. Меры однородности объектов. Расстояние между объектами. Меры близости между кластерами.
10. Методы кластерного анализа. Метод K-средних.
11. Предобработка признаков.
12. Отбор признаков
13. Метрики в задаче регрессии: MAE, MSE, MAPE, R 2.
14. Метрики в задаче классификации: кросс-энтропия, precision, recall, F-мера, ROC-кривая, AUC ROC.
15. Обобщающая способность модели и ее оценка: отложенная выборка, кросс-валидация.
16. Сущность главных компонент, их свойства, геометрическая интерпретация.
17. Области применения анализа данных.
18. Разбор типов данных и выбор подходящих под них графиков.
19. Работа с количественными и качественными данными.

20. Типы данных библиотеки numpy. Важнейшие стандартные функции.

21. Объект Series. Объект DataFrame.

22. Загрузка и выгрузка данных. Организация колонок и строчек. Пропуски и повторы.

23. Применение pandas в задаче исследования данных.

24. Виды графиков и диаграмм в *matplotlib, saebon, plotly*.

25. Основные элементы диаграммы в *matplotlib, saebon, plotly*.

26. Создание диаграммы в *matplotlib, saebon, plotly*.

27. График функции в *matplotlib, saebon, plotly*. Гистограмма в *matplotlib, saebon, plotly*.

28. Применение matplotlib, saebon, plotly для визуализации данных в процессе исследовательского анализа, решения задач классификации, регрессии, кластерного анализа и понижения размерности.

29. Визуализация моделей классификации, регрессии, кластерного анализа, результатов понижения размерности.

30. Визуальное представление оценок качества моделей классификации, регрессии, кластерного анализа, результатов понижения размерности.

31. Визуализация геопространственных данных.

32. Практика с онлайн и оффлайн сервисами визуализации данных.

Примерный перечень заданий к экзамену

Разработать приложение для решения следующей задачи:

1. Для заданного набора данных выполнить исследовательский анализ данных средствами pandas.

2. Для заданного набора данных выполнить кластерный анализ данных средствами scikit-learn.

3. Для заданного набора данных решить задачу классификации средствами scikit-learn.

4. Для заданного набора данных решить задачу понижения размерности данных средствами scikit-learn.

5. Визуализация моделей классификации, регрессии, кластерного анализа, результатов понижения размерности.

6. Визуальное представление оценок качества моделей классификации, регрессии, кластерного анализа, результатов понижения размерности.

7. Визуализация геопространственных данных.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Интеллектуальные информационные системы	Информацион-ных систем управления	Нет	Изменений не требуется (протокол № 18 от 08.06.2023 г.).

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры информационных систем управления (протокол № ____ от _____ 202_ г.)

Заведующий кафедрой

_____ _____ _____

(степень, звание) (подпись) (И.О.Фамилия)

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____

(степень, звание) (подпись) (И.О.Фамилия)