МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Экономический факультет

УТВЕРЖДАЮ:
Проректор по учебной работе,
качеству образования – первый
проректор

_____ Т.А. Хагуров
подпись

«26» мая 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.05 Data Mining

(код и наименование дисциплины в соответствии с учебным планом)

Направление подготовки: 27.03.03 Системный анализ и управление

(код и наименование направления подготовки/специальности)

Направленность (профиль):
Интеллектуальная бизнес-аналитика и управление экономическими процессами

(наименование направленности (профиля) / специализации)

Форма обучения: _____очная_____

(очная, очно-заочная, заочная)

Квалификация: бакалавр

Краснодар 2023

Рабочая программа дисциплины Б1.В.05 «Data Mining» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 27.03.03 «Системный анализ и управление»

Программу составил(и):

Н.Ю. Нарыжная, доцент кафедры экономики и управления инновационными системами, к.т.н., доцент

подпись

Рабочая программа дисциплины Б1.В.05 «Data Mining» утверждена на заседании кафедры экономики и управления инновационными системами протокол №5 «18» апреля 2023 г.
Заведующий кафедрой Литвинский К.О.

подпись

Утверждена на заседании учебно-методической комиссии экономического факультета протокол №8 «19» мая 2023 г.

Председатель УМК экономического факультета
Дробышевская Л.Н.

подпись

Рецензенты:

Гончаров В.А., и.о. директора ООО «АРТРЕ», г. Краснодар

Силинская С.М., к.т.н., доцент, доцент кафедры анализа данных и искусственного интеллекта ФГБОУ ВО «КубГУ»

1 Цели и задачи изучения дисциплины (модуля)

1.1 Цель освоения дисциплины

Цель освоения дисциплины «Data Mining» состоит в формировании знаний, умений и навыков (компетенций) по одному из приоритетных в современных информационных технологиях направлению – поиску и аналитической обработке данных.

1.2 Задачи дисциплины

– формирование у бакалавров практических навыков сбора, систематизации и хранения данных;

– формирование у бакалавров представления о технических и методологических средствах анализа данных, обеспечивающих сбор, хранение и систематизацию информации;

– ознакомление бакалавров с основными принципами машинного обучения - а именно, видами задач машинного обучения, классами моделей (линейные, логические, нейросетевые), метриками качествами и подходами к предварительной обработке данных.

1.3 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина Б1.В.05 «Data Mining» относится к части, формируемой участниками образовательных отношений Блока 1 «Дисциплины (модули)» учебного плана.

Перечень предшествующих дисциплин, необходимых для ее изучения:

- Информационно-коммуникационные технологии в профессиональной деятельности;
- Базы данных;
- Теория и технология программирования.

Перечень последующих дисциплин, для которых данная дисциплина является предшествующей в соответствии с учебным планом:

- Теория принятия решений;
- Технологическое предпринимательство;
- Анализ Big Data;
- Low-code аналитика;
- Аудит и оптимизация бизнес-процессов.

1.4 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих компетенций:

Код и наименование индикатора* достижения компетенции	Результаты обучения по дисциплине
ПК-3 Способен регламентировать процессы подразделений организации и разрабатывать административные регламенты подразделений организации (в том числе кросс-функциональные процессы)	
ИПК-3.1 Определяет источники, анализирует, собирает и систематизирует информацию для анализа	*Знает:* методы сбора информации (наблюдения, фиксация данных, хронометраж, фотография рабочего дня, техники проведения интервью и анкетирования, анализ документов и отчетной информации, изучение обратной связи от заинтересованных сторон) *Умеет:* выполнять наблюдения, интервью и анкетирование; анализировать, систематизировать и обобщать информацию; агрегировать, структурировать и обобщать информацию *Трудовое действие:* сбор информации о ходе и результатах процесса подразделения организации или административного регламента подразделения организации; оформление результатов сбора информации.
ПК-2 Способен анализировать и исследовать большие данные с использованием существующей в организации методологической и технологической инфраструктуры	

ИПК-2.2. Определяет источники, анализирует, собирает и систематизирует информацию для анализа	*Знает:* источники информации, в том числе информации, необходимой для обеспечения деятельности в предметной области заказчика исследования; виды источников данных: созданные человеком, созданные машинами; методы извлечения информации и знаний из гетерогенных, мультиструктурированных, неструктурированных источников, в том числе при потоковой обработке; режимы получения и обработки данных, поддержка режима реального времени. *Умеет:* осуществлять взаимодействие с внутренними и внешними поставщиками данных из гетерогенных источников; использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников, в том числе в режиме реального времени; производить очистку данных для проведения аналитических работ. *Трудовое действие:* извлечение, проверка и очистка больших объемов данных из гетерогенных источников; извлечение, проверка и очистка больших объемов данных из гетерогенных источников; оценка соответствия набора данных предметной области и задачам аналитических работ.
ПК-4 Способен обосновывать возможные решения и выбирать наиболее оптимальные	
ИПК-4.2 Определяет источники, анализирует, собирает и систематизирует информацию для анализа	*Знает*: методы сбора, анализа, систематизации, хранения и поддержания в актуальном состоянии информации бизнес-анализа. *Умеет:* проводить сбор и систематизацию информации; агрегировать, структурировать и обобщать; применять методы выбора оптимального решения. данные; готовить обоснование для выбора решения *Трудовое действие*: анализ, обоснование и выбор решения.

Результаты обучения по дисциплине достигаются в рамках осуществления всех видов контактной и самостоятельной работы обучающихся в соответствии с утвержденным учебным планом. Индикаторы достижения компетенций считаются сформированными при достижении соответствующих им результатов обучения.

2. Структура и содержание дисциплины

2.1 Распределение трудоёмкости дисциплины по видам работ

Общая трудоёмкость дисциплины составляет 3 зачетных единиц (108 часов), их распределение по видам работ представлено в таблице

Виды работ	Всего часов	Форма обучения			
		очная		очнозаочная	заочная
		5 семестр (часы)	X семестр (часы)	X семестр (часы)	X курс (часы)
Контактная работа, в том числе:	**58,2**	**58,2**			
Аудиторные занятия (всего):	**52**	**52**			
занятия лекционного типа	18	18			
лабораторные занятия	34	34			
практические занятия					
семинарские занятия					
Иная контактная работа:	**6,2**	**6,2**			
Контроль самостоятельной работы (КСР)	6	6			
Промежуточная аттестация (ИКР)	0,2	0,2			

Самостоятельная работа, в том числе:		**49,8**	**49,8**			
Расчётно-графическая работа (РГР) (подготовка)		15	15			
Самостоятельное изучение разделов, самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий, подготовка к лабораторным и практическим занятиям, коллоквиумам и т.д.)		34,8	34,8			
Контроль:						
Подготовка к зачету						
Общая трудоемкость	**час.**	**108**	**108**			
	в том числе контактная работа	**58,2**	**58,2**			
	зач. ед	**3**	**3**			

2.2 Содержание дисциплины

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.

Разделы (темы) дисциплины, изучаемые в 5 семестре *(очная форма обучения)*

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1.	Данные: источники хранения данных в сети, современные подходы к обработке	8	2		2	4
2.	Методы сбора и систематизации информации	10	2		4	4
3.	Сбор данных из социальных сетей. API протоколы для VK, LinkedIn, Twitter, Мой Мир.	10	2		4	4
4.	API протокол для финансовых данных: Marketstack, Finnhub, IEX Cloud, Finam.	10	2		4	4
5.	Обработка данных. Библиотеки NumPy, Pandas.	12	2		4	6
6.	Визуализация информации. Библиотеки Matplotlib и SeaBorn. Онлайн-сервисы для построения дашбордов.	14	2		6	6
7.	Парсинг сайтов на Python.	10	2		2	6
8.	Вэб-аналитика: Google Analytics, Яндекс.Метрика.	10	2		2	6
9.	Модели машинного обучения на данных: модели регрессии, классификации, кластеризации.	17,8	2		6	9,8
	ИТОГО по разделам дисциплины	101, 8	18		34	49,8
	Контроль самостоятельной работы (КСР)	6				
	Промежуточная аттестация (ИКР)	0,2				
	Подготовка к текущему контролю					
	Общая трудоемкость по дисциплине	**108**				

Примечание: Л - лекции, ПЗ - практические занятия / семинары, ЛР - лабораторные занятия, СРС - самостоятельная работа студента

2.3 Содержание разделов (тем) дисциплины

2.3.1 Занятия лекционного типа

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1.	Данные: источники хранения информации в сети, современные подходы к обработке	Анализ источников информации в сети Интернет. Классификация данных.	*Опрос*
2.	Методы сбора и систематизации информации	Этапы исследования. Типы необходимой для исследования информации. Виды исследований. Методы сбора данных. Инструменты для сбора данных.	*Опрос*
3.	Сбор данных из социальных сетей. API протокол для VK, LinkedIn, Twitter, Мой Мир.	API как средство интеграции приложений. Типы API-интерфейсов. Подключение к Twitter API. Исследование шаблонов в ретвитах. Определения лексического разнообразия твитов. Выполнение запросов к LinkedIn. Загрузка файла с информацией о контактах с LinkedIn. Нормализация данных для анализа.	*Опрос*
4.	API протоколы для финансовых данных: Marketstack, Finnhub, IEX Cloud, Finam.	Интерфейс API REST для получения данных фондового рынка. API фондового рынка Finnhub Stock API для построения финансовых продуктов. Способы получения и использования точных рыночных данных с помощью IEX Cloud API.	*Опрос*
5.	Обработка данных. Библиотеки NumPy, Pandas.	Программный минимум для анализа данных: NumPy и Pandas. Структуры данных Series и DataFrame. Первичный анализ данных с Pandas, демонстрация основных методов Pandas и NumPy. Задача	*Тест*
6.	Визуализация информации. Библиотеки Matplotlib и SeaBorn. Онлайн-сервисы для построения дашбордов.	Демонстрация основных методов Seaborn. Построение bar chart, pair plot, dist plot, joint plot, box plot. Интерактивные графики в jupyter notebook с помощью библиотеки Plotly. Визуализация задачи оттока клиентов.	*Опрос*
7.	Парсинг сайтов на Python.	Автоматизированный сбор информации с сайтов. Технический парсинг сайта. Анализ структуры сайтов-конкурентов с целью улучшения и развития собственной структуры. Задача парсинга названий товаров, артикулов, цен для наполнения своего собственного интернет- магазина.	*Опрос*
8.	Вэб-аналитика: Google Analytics, Яндекс.Метрика.	Статистика посещаемости разделов и веб-страниц сайта: количество просмотренных веб-страниц, ключевые слова и фразы, по которым посетители находят сайт в поисковых системах, география посетителей, время, проведенное на веб-странице посетителем. Методы и инструменты вебаналитики. Анализ поведения посетителей на странице: взаимодействие с формами, совершение микро и макроконверсий. Сквозная аналитика. Отслеживание полного пути пользователя от просмотра рекламы и до завершения сделки, а также повторных продаж.	*Опрос*
9.	Модели машинного обучения на данных: модели регрессии, классификации, кластеризации.	Типы задач машинного обучения. Введение. Постановки задач в машинном обучении. Обучение с учителем и без. Классификация, регрессия, ранжирование, кластеризация. Обучающая и тестовая выборки. Проблема переобучения. Кросс-валидация. Метод ближайших соседей. Решающие деревья. Бэггинг и метод случайных подпространств. Случайные леса. Бустинг. Градиентный бустинг над решающими деревьями. Модель xgboost.	*Опрос, Тест*

2.3.2 Занятия семинарского типа (практические / семинарские занятия/ лабораторные работы)

№	Наименование раздела (темы)	Тематика занятий/работ	Форма текущего контроля
1.	Данные: источники хранения информации в сети, современные подходы к обработке;	Поиск данных. Хранение больших данных	*ЛР*
2.	Методы сбора и систематизации информации	Этапы исследования. Типы необходимой для исследования информации. Виды исследований. Методы сбора данных. Инструменты для сбора данных.	*ЛР*
3.	Сбор данных из социальных сетей. API протоколы для VK, LinkedIn, Twitter, Мой Мир.	Программный интерфейс API reference index. Инструмент командной строки Twurl	*ЛР, Т*
4.	API протокол для финансовых данных: Marketstack, Finnhub, IEX Cloud, Finam.	Получение рыночных данных с помощью стандартных API-интерфейсов	*ЛР*
5.	Обработка данных. Библиотеки NumPy, Pandas.	Загрузка, запрос и фильтрация данных с помощью библиотеки Pandas.	*ЛР, Т*
6.	Визуализация информации. Библиотеки Matplotlib и SeaBorn. Онлайн-сервисы для построения дашбордов.	Демонстрация основных методов Seaborn. Построение bar chart, pair plot, dist plot, joint plot, box plot. Интерактивные графики в jupyter notebook с помощью библиотеки Plotly. Визуализация задачи оттока клиентов.	*ЛР*
7.	Парсинг сайтов на Python.	Автоматизированный сбор информации с сайтов. Технический парсинг сайта. Анализ структуры сайтов-конкурентов с целью улучшения и развития собственной структуры. Задача парсинга названий товаров, артикулов, цен для наполнения своего собственного интернет-магазина.	*ЛР*
8.	Вэб-аналитика: Google Analytics, Яндекс.Метрика.	Системы интернет-аналитики с детализацией поведения посетителя на странице. Google Analytics, Яндекс.Метрика	*ЛР*
9.	Модели машинного обучения на данных: модели регрессии, классификации, кластеризации.	Построение и отбор признаков. Приложения в задачах обработки текста, изображений и геоданных. Обучение без учителя: PCA, кластеризация.	*ЛР*

Защита лабораторной работы (ЛР), выполнение курсового проекта (КП), курсовой работы (КР), расчетно-графического задания (РГЗ), написание реферата (Р), эссе (Э), коллоквиум (К), тестирование (Т) и т.д.

2.3.3 Примерная тематика курсовых работ (проектов)

Не предусмотрено

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	Занятия лекционного и семинарского типа	Методические указания для подготовки к занятиям лекционного и семинарского типа. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа:
2	Выполнение самостоятельной работы обучающихся	Методические указания по выполнению самостоятельной работы обучающихся. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
3	Выполнение расчетно-графических заданий	Методические указания по выполнению расчетно-графических заданий. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
4	Выполнение лабораторных работ	Методические указания по выполнению лабораторных работ. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
10	Интерактивные методы обучения	Методические указания по интерактивным методам обучения. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:
- в печатной форме увеличенным шрифтом,
- в форме электронного документа,
- в форме аудиофайла,
- в печатной форме на языке Брайля.

Для лиц с нарушениями слуха:
- в печатной форме,
- в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:
- в печатной форме,
- в форме электронного документа,
- в форме аудиофайла.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. Образовательные технологии, применяемые при освоении дисциплины (модуля)

В ходе изучения дисциплины предусмотрено использование следующих образовательных технологий: лекции, лабораторные занятия, подготовка письменных расчетно-графических работ, самостоятельная работа студентов.

Компетентностный подход в рамках преподавания дисциплины реализуется в использовании интерактивных технологий и активных методов (проектных методик, мозгового штурма, разбора конкретных ситуаций) в сочетании с внеаудиторной работой.

Информационные технологии, применяемые при изучении дисциплины: использование информационных ресурсов, доступных в информационно-телекоммуникационной сети Интернет.

Адаптивные образовательные технологии, применяемые при изучении дисциплины - для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

4. Оценочные средства для текущего контроля успеваемости и промежуточной аттестации

Оценочные средства предназначены для контроля и оценки образовательных достижений обучающихся, освоивших программу учебной дисциплины Сбор и систематизация информации.

Оценочные средства включает контрольные материалы для проведения **текущего контроля** в форме *опроса, расчетно-графических заданий лабораторных работ* и **промежуточной аттестации** в форме вопросов и заданий к зачету.

Структура оценочных средств для текущей и промежуточной аттестации

№ п/п	Код и наименование индикатора (в соответствии с п. 1.4)	Результаты обучения (в соответствии с п. 1.4)	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
1	ИПК-3.1 Определяет источники, анализирует, собирает и систематизирует информацию для анализа	Знает методы сбора информации (наблюдения, фиксация данных, хронометраж, фотография рабочего дня, техники проведения интервью и анкетирования, анализ документов и отчетной информации, изучение обратной связи от заинтересованных сторон)	Опрос	ЛР Вопросы на зачете
		Умеет: выполнять наблюдения, интервью и анкетирование	РГЗ	ЛР Вопросы на зачете
		Умеет: анализировать, систематизировать и обобщать информацию	РГЗ	ЛР Вопросы на зачете
		Умеет: агрегировать, структурировать и обобщать информацию	Опрос	ЛР Вопросы на зачете
		Трудовое действие: сбор информации о ходе и результатах процесса подразделения организации или административного регламента подразделения организации	Опрос	ЛР Вопросы на зачете
		Трудовое действие: оформление результатов сбора информации	РГЗ	ЛР Вопросы на зачете

2	ИПК-2.2. Определяет источники, анализирует, собирает и систематизирует информацию для анализа	*Знает:* источники информации, в том числе информации, необходимой для обеспечения деятельности в предметной области заказчика исследования; виды источников данных: созданные человеком, созданные машинами	Опрос	ЛР Вопросы на зачете
		Знает: методы извлечения информации и знаний из гетерогенных, мультиструктурированных, неструктурированных источников, в том числе при потоковой обработке; режимы получения и обработки данных, поддержка режима реального времени	Опрос	ЛР Вопросы на зачете
		Умеет: осуществлять взаимодействие с внутренними и внешними поставщиками данных из гетерогенных источников	РГЗ	ЛР Вопросы на зачете
		Умеет: использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников, в том числе в режиме реального времени; производить очистку данных для проведения аналитических работ	Опрос, РГЗ	ЛР Вопросы на зачете
		Трудовое действие: извлечение, проверка и очистка больших объемов данных из гетерогенных источников	РГЗ	ЛР Вопросы на зачете
		Трудовое действие: извлечение, проверка и очистка больших объемов данных из гетерогенных источников	Опрос	ЛР Вопросы на зачете
		Трудовое действие: оценка соответствия набора данных предметной области и задачам аналитических работ	Опрос	ЛР Вопросы на зачете
3	ИПК-4.2 Определяет источники, анализирует, собирает и систематизирует информацию для анализа	*Знает:* методы сбора, анализа, систематизации, хранения и поддержания в актуальном состоянии информации бизнес-анализа	Опрос	ЛР Вопросы на зачете

		Умеет: проводить сбор и систематизацию информации	Опрос	ЛР Вопросы на зачете
		Умеет: агрегировать, структурировать и обобщать данные	РГЗ	ЛР Вопросы на зачете
		Умеет: применять методы выбора оптимального решения	РГЗ	ЛР Вопросы на зачете
		Трудовое действие: анализ, обоснование и выбор решения	РГЗ	ЛР Вопросы на зачете

Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Расчетно-графическое задание на тему: «Метод главных компонент (PCA), кластеризация»

Мы будем работать с набором данных Samsung Human Activity Recognition. Скачайте данные отсюда. Данные поступают с акселерометров и гироскопов мобильных телефонов Samsung Galaxy S3 (подробнее про признаки - по ссылке на UCI выше), также известен вид активности человека с телефоном в кармане - ходил ли он, стоял, лежал, сидел или шел вверх/вниз по лестнице.

Вначале мы представим, что вид активности нам неизвестнен, и попробуем кластеризовать людей просто на основе имеющихся признаков. Затем решим задачу определения вида физической активности именно как задачу классификации.

Заполните код в клетках (где написано "Ваш код здесь") и ответьте на вопросы.

Подготовительный этап (загрузка библиотек и стилей)

```
import numpy as np
import pandas as pd
import seaborn as sns
from tqdm import tqdm notebook

%matplotlib inline
from matplotlib import pyplot as plt

plt.style.use(['seaborn-darkgrid'])
plt.rcParams['figure.figsize'] = (12, 9)
plt.rcParams['font.family'] = 'DejaVu Sans'

from sklearn import metrics
from sklearn.cluster import AgglomerativeClustering, KMeans, SpectralClusteri
ng
from sklearn.decomposition import PCA
from sklearn.model_selection import GridSearchCV

from  sklearn.preprocessing import StandardScaler
```

```
from  sklearn.svm import LinearSVC

RANDOM_STATE =  17
X train = np.loadtxt("../../data/samsung HAR/samsung train.txt")
y_train = np.loadtxt("../../data/samsung_HAR/samsung_train_labels.txt").astyp
e(int)

X_test = np.loadtxt("../../data/samsung_HAR/samsung_test.txt")
y_test = np.loadtxt("../../data/samsung_HAR/samsung_test_labels.txt").astype(
int)
# Проверим размерности
assert(X train.shape ==  (7352, 561) and y train.shape ==  (7352,))
assert(X test.shape ==  (2947, 561) and y test.shape == (2947,))
```

Для кластеризации нам не нужен вектор ответов, поэтому будем работать с объединением обучающей и тестовой выборок. Объедините *Xtrain* с *X_test*, а *ytrain* - с *ytest.*

```
# Ваш код здесь
```

X = y = Определим число уникальных значений меток целевого класса.

```
np.unique(y) array([1, 2,     3,    4,    5,    6])
n classes = np.unique(y).size
```

Отмасштабируйте выборку с помощью StandardScaler с параметрами по умолчанию.

```
# Ваш код здесь scaler =
X_scaled =
```

Понижаем размерность с помощью PCA, оставляя столько компонент, сколько нужно для того, чтобы объяснить как минимум 90% дисперсии исходных (отмасштабированных) данных. Используйте отмасштабированную выборку и зафиксируйте random_state (константа RANDOM_STATE).

```
# Ваш код здесь pca =
X_pca =
```

Вопрос 1:
Какое минимальное число главных компонент нужно выделить, чтобы объяснить 90% дисперсии исходных (отмасштабированных) данных?

```
# Ваш код здесь
```

Варианты:

- 56
- 65
- 66
- 193

Вопрос 2:
Сколько процентов дисперсии приходится на первую главную компоненту? Округлите до целых процентов.

Варианты:

- 45
- 51

- 56
- 61

```
#  Ваш код здесь
```

Визуализируйте данные в проекции на первые две главные компоненты.

```
#  Ваш код здесь
plt.scatter(, , c=y, s=20, cmap='viridis');
```

Вопрос 3:
Если все получилось правильно, Вы увидите сколько-то кластеров, почти идеально отделенных друг от друга. Какие виды активности входят в эти кластеры?

Ответ:

- 1 кластер: все 6 активностей
- 2 кластера: (ходьба, подъем вверх по лестнице, спуск по лестнице) и (сидение, стояние, лежание)
- 3 кластера: (ходьба), (подъем вверх по лестнице, спуск по лестнице) и (сидение, стояние, лежание)
- 6 кластеров

Сделайте кластеризацию данных методом KMeans, обучив модель на данных со сниженной за счет PCA размерностью. В данном случае мы подскажем, что нужно искать именно 6 кластеров, но в общем случае мы не будем знать, сколько кластеров надо искать.

Параметры:

- **n_clusters** = n_classes (число уникальных меток целевого класса)
- **n_init** = 100
- **random_state** = RANDOM_STATE (для воспроизводимости результата)

Остальные параметры со значениями по умолчанию.

```
#  Ваш код здесь
```

Визуализируйте данные в проекции на первые две главные компоненты. Раскрасьте точки в соответствии с полученными метками кластеров.

```
#  Ваш код здесь
plt.scatter(, , c=cluster labels, s=20,  cmap='viridis');
```

Посмотрите на соответствие между метками кластеров и исходными метками классов и на то, какие виды активностей алгоритм KMeans путает.

```
tab = pd.crosstab(y, cluster labels, margins=True)
tab.index =  ['ходьба', 'подъем вверх по лестнице',
              'спуск по лестнице', 'сидение', 'стояние', 'лежание', 'все']
tab.columns =  ['cluster' + str(i + 1) for i in range(6)] +  ['все']

tab
```

Видим, что каждому классу (т.е. каждой активности) соответствуют несколько кластеров. Давайте посмотрим на максимальную долю объектов в классе, отнесенных к какому-то одному кластеру. Это будет простой метрикой, характеризующей, насколько легко класс отделяется от других при кластеризации.

Пример: если для класса "спуск по лестнице", в котором 1406 объектов, распределение кластеров такое:

- кластер 1 - 900
- кластер 3 - 500
- кластер 6 - 6,

то такая доля будет 900 / 1406 ~~ 0.64.

Вопрос 4:
Какой вид активности отделился от остальных лучше всего в терминах простой метрики, описанной выше?

Ответ:

- ходьба
- стояние
- спуск полестнице
- перечисленные варианты не подходят

Видно, что kMeans не очень хорошо отличает только активности друг от друга.
Используйте метод локтя, чтобы выбрать оптимальное количество кластеров. Параметры алгоритма и данные используем те же, что раньше, меняем только n_clusters.

```
# Ваш код здесь
inertia = []
for k in tqdm_notebook(range(1, n_classes + 1)):
    #
    #
```

Вопрос 5:
Какое количество кластеров оптимально выбрать, согласно методу локтя?

Ответ:

- 1
- 2
- 3
- 4

Попробуем еще один метод кластеризации, который описывался в статье - агломеративную кластеризацию.

```
ag = AgglomerativeClustering(n_clusters=n_classes, linkage='ward').fit(X_pca)
```

Посчитайте Adjusted Rand Index (sklearn.metrics) для получившегося разбиения на кластеры и для KMeans с параметрами из задания к 4 вопросу.

```
# Ваш код здесь
```

Вопрос 6:
Отметьте все верные утверждения.

Варианты:

- Согласно ARI, KMeans справился с кластеризацией хуже, чем Agglomerative Clustering
- Для ARI не имеет значения какие именно метки присвоены кластерам, имеет значение только разбиение объектов на кластеры
- В случае случайного разбиения на кластеры ARI будет близок к нулю

Можно заметить, что задача не очень хорошо решается именно как задача кластеризации, если выделять несколько кластеров (> 2). Давайте теперь решим задачу классификации, вспомнив, что данные у нас размечены.

Для классификации используйте метод опорных векторов - класс sklearn.svm.LinearSVC.

Настройте для LinearSVC гиперпараметр C с помощью GridSearchCV.

- Обучите новый StandardScaler на обучающей выборке (со всеми исходными признаками), прмиените масштабирование к тестовой выборке
- В GridSearchCV укажите cv=3.

```
# Ваш код здесь
#
X train scaled =
X test scaled =
svc = LinearSVC(random state=RANDOM STATE) svc params =    {'C': [0.001,
             0.01, 0.1,    1,    10]}
# Ваш код здесь best_svc =
# Ваш код здесь
```

Вопрос 7
Какое значение гиперпараметра C было выбрано лучшим по итогам кросс-валидации?

Ответ:

- 0.001
- 0.01
- 0.1
- 1
- 10

```
y_predicted = best_svc.predict(X_test_scaled)
tab = pd.crosstab(y test, y predicted, margins=True)
tab.index =  ['ходьба', 'подъем вверх по лестнице', 'спуск по лестнице',
              'сидение', 'стояние', 'лежание', 'все']
tab.columns = tab.index tab
```

Вопрос 8:
Какой вид активности SVM определяет хуже всего в терминах точности? Полноты?

Ответ:

- по точности - подъем вверх по лестнице, по полноте - лежание
- по точности - лежание, по полноте - сидение
- по точности - ходьба, по полноте - ходьба
- по точности - стояние, по полноте - сидение

Наконец, проделайте то же самое, что в 7 вопросе, только добавив PCA.

- Используйте выборки X_train_scaled и X_test_scaled
- Обучите тот же PCA, что раньше, на отмасшабированной обучающей выборке,

примените преобразование к тестовой
- Настройте гиперпараметр C на кросс-валидации по обучающей выборке с PCA-преобразованием. Вы заметите, насколько это проходит быстрее, чем раньше.

Вопрос 9:
Какова разность между лучшим качеством (долей верных ответов) на кросс-валидации в случае всех 561 исходных признаков и во втором случае, когда применялся метод главных компонент? Округлите до целых процентов.

Варианты:

- Качество одинаковое
- 2%
- 4%
- 10%
- 20%

Вопрос 10:
Выберите все верные утверждения:

Варианты:

- Метод главных компонент в данном случае позволил уменьшить время обучения модели, при этом качество (доля верных ответов на кросс-валидации) очень пострадало, более чем на 10%
- PCA можно использовать для визуализации данных, однако для этой задачи есть и лучше подходящие методы, например, tSNE. Зато PCA имеет меньшую вычислительную сложность
- PCA строит линейные комбинации исходных признаков, и в некоторых задачах они могут плохо интерпретироваться человеком

Зачетно-экзаменационные материалы для промежуточной аттестации (зачет)

1. Источники хранения данных в сети Интернет. Классификация данных.
2. Основные подходы обработки данных.
3. Сбор данных из социальных сетей.
4. API протокол для VK, LinkedIn
5. API протокол для Твиттер.
6. API протоколы для финансовых данных.
7. API фондового рынка Finnhub Stock
8. Библиотека анализа данных NumPy
9. Библиотека Pandas.
10. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
11. Метрические методы классификации. Метрики качества алгоритмов классификации и регрессии.
12. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.
13. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.
14. Деревья решений. Методы построения деревьев.
15. Случайный лес, его особенности.
16. Градиентный бустинг, его особенности при использовании деревьев в качестве

базовых алгоритмов.

17. Визуализация информации. Библиотеки Matplotlib и SeaBorn.
18. Автоматизированный сбор информации с сайтов. Технический парсинг сайта.
19. Обучение без учителя. Методы понижения размерности. PCA.
20. Обучение без учителя. Кластеризация.
21. Вэб-аналитика: Google Analytics, Яндекс.Метрика.

Критерии оценивания результатов обучения

Оценка	Критерии оценивания по зачету
зачтено	Высокий уровень (студент, освоивший знания, умения, компетенции и теоретический материал без пробелов; выполнивший все задания, предусмотренные учебным планом на высоком качественном уровне; практические навыки профессионального применения освоенных знаний сформированы).
	Средний уровень (студент, практически полностью освоивший знания, умения, компетенции и теоретический материал, учебные задания не оценены максимальным числом баллов, в основном сформировал практические навыки).
	Пороговый уровень (студент, частично с пробелами освоивший знания, умения, компетенции и теоретический материал, многие учебные задания либо не выполнил, либо они оценены числом баллов близким к минимальному, некоторые практические навыки не сформированы).
не зачтено	Минимальный уровень (студент, не освоивший знания, умения, компетенции и теоретический материал, учебные задания не выполнил, практические навыки не сформированы).

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

- при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;
- при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;
- при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:
- в печатной форме увеличенным шрифтом,
- в форме электронного документа.

Для лиц с нарушениями слуха:
- в печатной форме,
- в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:
- в печатной форме,
- в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента

обучающихся.

2. Перечень учебной литературы, информационных ресурсов и технологий

2.1. Учебная литература

1. Анализ данных : учебник для академического бакалавриата / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2020. — 490 с. — (Бакалавр. Академический курс). — ISBN 978-5-534-00616-2. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/412967

2. Платонов**,** А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — Москва : Издательство Юрайт, 2022. — 85 с. — (Высшее образование). — ISBN 978-5534-15561-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/508804.

3. Миркин*, Б. Г.* Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2022. — 174 с. — (Высшее образование). — ISBN 978-59916-5009-0. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/469306 .

2.2. Периодическая литература

Указываются печатные периодические издания из «Перечня печатных периодических изданий, хранящихся в фонде Научной библиотеки КубГУ» https://www.kubsu.ru/ru/node/15554, и/или электронные периодические издания, с указанием адреса сайта электронной версии журнала, из баз данных, доступ к которым имеет КубГУ:

1. Базы данных компании «Ист Вью» http://dlib.eastview.com
2. Электронная библиотека GREBENNIKON.RU https://grebennikon.ru/

2.3. Интернет-ресурсы, в том числе современные профессиональные базы данных и информационные справочные системы

Электронно-библиотечные системы (ЭБС):

1. ЭБС «ЮРАЙТ» https://urait.ru/
2. ЭБС «УНИВЕРСИТЕТСКАЯ БИБЛИОТЕКА ОНЛАЙН» www.biblioclub.ru
3. ЭБС «BOOK.ru» https://www.book.ru
4. ЭБС «ZNANIUM.COM» www.znanium.com
5. ЭБС «ЛАНЬ» https://edanbook.com

Профессиональные базы данных:

1. Scopus http://www.scopus.com/
2. ScienceDirect www.sciencedirect.com
3. Журналы издательства Wiley https://onlinelibrary.wiley.com/
4. Научная электронная библиотека (НЭБ) http://www.elibrary.ru/
5. Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН http://archive.neicon.ru
6. Национальная электронная библиотека (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) https://rusneb.ru/
7. Президентская библиотека им. Б.Н. Ельцина https://www.prlib.ru/
8. Электронная коллекция Оксфордского Российского Фонда https://ebookcentral.proquest.com/lib/kubanstate/home.action
9. Springer Journals https://link.springer.com/
10. Nature Journals https://www.nature.com/siteindex/index.html
11. Springer Nature Protocols and Methods https://experiments.springernature.com/sources/springer-protocols
12. Springer Materials http://materials.springer.com/
13. zbMath https://zbmath.org/
14. Nano Database https://nano.nature.com/

15. Springer eBooks: https://link.springer.com/
16. "Лекториум ТВ" http://www.lektorium.tv/
17. Университетская информационная система РОССИЯ http://uisrussia.msu.ru

Информационные справочные системы:

1. Консультант Плюс - справочная правовая система (доступ по локальной сети с компьютеров библиотеки)

Ресурсы свободного доступа:

1. Американская патентная база данных http://www.uspto.gov/patft/
2. КиберЛенинка (http://cyberleninka.ru/);
3. Министерство науки и высшего образования Российской Федерации https://www.minobrnauki.gov.ru/;
4. Федеральный портал "Российское образование" http://www.edu.ru/;
5. Информационная система "Единое окно доступа к образовательным ресурсам" http://window.edu.ru/;
6. Единая коллекция цифровых образовательных ресурсов http://school-collection.edu.ru/ .
7. Проект Государственного института русского языка имени А.С. Пушкина "Образование на русском" https://pushkininstitute.ru/;
8. Справочно-информационный портал "Русский язык" http://gramota.ru/;
9. Служба тематических толковых словарей http://www.glossary.ru/;
10. Словари и энциклопедии http://dic.academic.ru/;
11. Образовательный портал "Учеба" http://www.ucheba.com/;
12. Законопроект "Об образовании в Российской Федерации". Вопросы и ответы http://xn--273--84d1f.xn--p1ai/voprosy_i_otvety

Собственные электронные образовательные и информационные ресурсы КубГУ:

1. Электронный каталог Научной библиотеки КубГУ http://megapro.kubsu.ru/MegaPro/W eb

2. Электронная библиотека трудов ученых КубГУ http://megapro.kubsu.ru/MegaPro/UserEntry?Action=ToDb&idb=6
3. Среда модульного динамического обучения http://moodle.kubsu.ru
4. База учебных планов, учебно-методических комплексов, публикаций и конференций http://mschool.kubsu.ru/
5. Библиотека информационных ресурсов кафедры информационных образовательных технологий http://mschool.kubsu.ru;
6. Электронный архив документов КубГУ http://docspace.kubsu.ru/
7. Электронные образовательные ресурсы кафедры информационных систем и технологий в образовании КубГУ и научно-методического журнала "ШКОЛЬНЫЕ ГОДЫ" http://icdau.kubsu.ru/

6. Методические указания для обучающихся по освоению дисциплины (модуля)

Самостоятельная работа студентов является неотъемлемой частью процесса подготовки. Дисциплину рекомендуется изучать путем систематической проработки лекционного материала, самостоятельной проработки рекомендуемой литературы, руководств и методических указаний к выполнению практических занятий. Цель самостоятельной работы - расширение кругозора и углубление знаний в области финансового инструментария.

Контроль за выполнением самостоятельной работы проводится при изучении каждой темы дисциплины на семинарских занятиях. Это текущий опрос, тестовые задания, контрольная работа.

В часы, отведенные для самостоятельной работы, студенты под руководством преподавателя обязаны выполнять индивидуальные практические задания, полученные на практических занятиях. При выполнении этих заданий необходимо использовать теоретический материал, делать ссылки на соответствующие формулы, проверять выполнимость предпосылок, необходимых для применения того или иного метода.

В освоении дисциплины инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная учебная работа (консультации) - дополнительное разъяснение учебного материала.

Индивидуальные консультации по предмету являются важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или лицом с ограниченными возможностями здоровья.

7. Материально-техническое обеспечение по дисциплине (модулю)

По всем видам учебной деятельности в рамках дисциплины используются аудитории, кабинеты и лаборатории, оснащенные необходимым специализированным и лабораторным оборудованием.

Наименование специальных помещений	Оснащенность специальных помещений	Перечень лицензионного программного обеспечения
Учебные аудитории для проведения занятий лекционного типа	Мебель: учебная мебель Технические средства обучения: экран, проектор, ноутбук	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Учебные аудитории для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации	Мебель: учебная мебель Технические средства обучения: экран, проектор, ноутбук	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Учебные аудитории для проведения лабораторных работ	Мебель: учебная мебель Технические средства обучения: экран, проектор, компьютеры, ноутбуки Оборудование:	

Лаборатория информационных и управляющих систем 201Н Лаборатория экономической информатики 202Н Лаборатория управления в технических системах 207Н	ПК, Терминальные станции, Усилитель автономный беспроводной Типовой комплект учебного оборудования "Теория автоматического управления", Презентации и плакаты Усилитель автономный беспроводной с микрофоном	Microsoft Windows 8, 10, Microsoft Office Professional Plus 1С: Предприятие 8 SPSS Statistics Microsoft Windows 8, 10, Microsoft Office Professional Plus
Лаборатория организационно-технологического обеспечения торговой и маркетинговой деятельности 201А	Панель интерактивная, Конференц-система, Микшер- усилитель, Подавитель акустической обратной связи, Настенный громкоговоритель, Радиосистема, Микрофон на гибком держателе, Моноблок HP, Документ-камера, Беспроводная точка доступа, Система видеоотображения, ЖК панель, Сплитер, Мультимедийная трибуна лектор, Система видеоконференцсвязи, Плакаты	Microsoft Windows 8, 10, Microsoft Office Professional Plus 1С: Предприятие 8
Лаборатория экономики и управления 212Н	Презентации и плакаты, Многофункциональный профессиональный видео детектор банкнот и ценных бумаг, Счетчики банкнот, Инфракрасный детектор банкнот и ценных бумаг, Универсальный детектор банкнот и ценных бумаг, Детектор подлинности банкнот, Ящик денежный, Планшетный импринтер, Усилитель автономный беспроводной	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Лаборатория безопасности жизнедеятельности 105А	Лабораторные стенды, Типовой комплект учебного оборудования, Стенды- тренажеры, Стенд-планшет, Тренажерный комплекс по применению первичных средств пожаротушения, Комплекс - тренажер по оказанию первой доврачебной помощи, Робот-тренажер, Комплект плакатов, Комплект демонстрационных пособий, Комплект аудиовизуальных пособий	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Учебные аудитории для курсового проектирования (выполнения курсовых работ)	Мебель: учебная мебель Технические средства обучения: экран, проектор, компьютер	Microsoft Windows 8, 10, Microsoft Office Professional Plus

Для самостоятельной работы обучающихся предусмотрены помещения, укомплектованные специализированной мебелью, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.

Наименование помещений для самостоятельной работы обучающихся	Оснащенность помещений для самостоятельной работы обучающихся	Перечень лицензионного программного обеспечения
Помещение для самостоятельной работы обучающихся (читальный зал Научной библиотеки)	Мебель: учебная мебель Комплект специализированной мебели: компьютерные столы Оборудование: компьютерная техника с подключением к информацио нно - коммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду образовательной организации, веб-камеры, коммуникационное оборудование, обеспечивающее доступ к сети интернет (проводное соединение и беспроводное соединение по технологии Wi-Fi)	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Помещение для самостоятельной работы обучающихся (ауд.213 А, 218 А)	Мебель: учебная мебель Комплект специализированной мебели: компьютерные столы Оборудование: компьютерная техника с подключением к информационнокоммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду образовательной организации, веб-камеры, коммуникационное оборудование, обеспечивающее доступ к сети интернет (проводное соединение и беспроводное соединение по технологии Wi-Fi)	Microsoft Windows 8, 10, Microsoft Office Professional Plus