МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ
_____

_____2024

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

дисциплины: **Технологии интеллектуального анализа BigData в экономических исследованиях**
специальность: **38.05.01 Экономическая безопасность**
специализация: **Экономическая безопасность бизнеса в цифровой экономике**
форма обучения: **очная**

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 2026062FC6FB7E4F1F37670D47CAB35B
Владелец: Клочков Юрий Сергеевич
Действителен: с 09.02.2024 до 04.05.2025

Фонд оценочных средств разработан по специальности 38.05.01 Экономическая безопасность, специализация «Экономическая безопасность бизнеса в цифровой экономике».

1. Формы аттестации по дисциплине

1.1. Форма промежуточной аттестации: ***зачет.***

Способ проведения промежуточной аттестации: ***устный зачет***

1.2.Формы текущей аттестации:

Таблица 1.1

№ п/п	Форма обучения
	ОФО
1	Тестирование Практическое задание
2	Тестирование Практическое задание
3	Тестирование Практическое задание
4	Тестирование Практическое задание

2. Результаты обучения по дисциплине, подлежащие проверке при проведении текущей и промежуточной аттестации

Таблица 2.1

№ п/п	Структурные элементы дисциплины		Код результата обучения по дисциплине	Оценочные средства	
	Номер раздела	Дидактические единицы (предметные темы)		Текущая аттестация	Промежуточная аттестация
1	1	Введение в интеллектуальный анализ данных	З1; У1; В1 З2; У2; В2 З3; У3; В3 З4; У4; В4	Тест практическое задание	Перечень вопросов к зачету
2	2	Интеллектуальный анализ данных, извлечение знаний из данных	З1; У1; В1 З2; У2; В2 З3; У3; В3 З4; У4; В4	Тест практическое задание	
3	3	Регрессионный анализ. Кластеризация	З1; У1; В1 З2; У2; В2 З3; У3; В3 З4; У4; В4	Тест практическое задание	
4	4	Визуальный анализ данных	З1; У1; В1 З2; У2; В2 З3; У3; В3 З4; У4; В4	Тест практическое задание	

3. Фонд оценочных средств

3.1. Фонд оценочных средств, позволяющие оценить результаты обучения по дисциплине, включает в себя оценочные средства для текущей аттестации.

3.2. Фонд оценочных средств для текущей аттестации включает:

- комплект тестов по разделу: «Введение в интеллектуальный анализ данных» - 10 шт. (Приложение 1);
- комплект тестов по разделу: «Интеллектуальный анализ данных, извлечение знаний из данных» - 10 шт. (Приложение 2);

- комплект тестов по разделу: «Регрессионный анализ. Кластеризация» - 10 шт. Приложение 3);

- комплект тестов по разделу: «Визуальный анализ данных» - 10 шт. (Приложение 4);

- комплект практических задания по разделу: «Введение в интеллектуальный анализ данных» - 1 шт. (Приложение 5);

- комплект практических задания по разделу: «Интеллектуальный анализ данных, извлечение знаний из данных» - 1 шт. (Приложение 6);

- комплект практических задания по разделу: «Регрессионный анализ. Кластеризация» - 1 шт. (Приложение 7);

- комплект практических задания по разделу: «Визуальный анализ данных» - 1 шт. (Приложение 8).

3.3. Фонд оценочных средств для промежуточной аттестации включает:

- комплект вопросов для зачета - 20 шт. (Приложение 9)

Приложение 1

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Комплект тестов
к разделу «Введение в интеллектуальный анализ данных»

1. Что такое бизнес-процесс?
a) Любая деятельность в корпоративных масштабах
b) Коммерческая деятельность с целью получения прибыли
c) Совокупность бизнес-функций
d) Последовательность действий по преобразованию входов в выходы, удовлетворяющие потребителя

2. Описать структуру системы бизнес-процессов, показать состав процессов одного уровня абстракции и взаимосвязи между ними можно с помощью диаграммы в нотации
a) EPC
b) IDEF0
c) BPMN
d) DFD

3. Архитектура предприятия — это
a) Искусство проектировать и строить бизнес-центры и производственные здания
b) Концептуальная структура организация системы
c) Единая система, которая описывает существующие организационные структуры, цели и показатели их достижения, линейку создаваемых продуктов/услуг, которые приносят доход, а также инфраструктуру (программное и аппаратное обеспечение, оборудование), используемые в работе
d) Стиль управления

4. Требование «Пользовательский GUI должен предоставлять возможность языковой локализации: выбор языка (русский/английский) для надписей на элементах» — это
a) Требование стейкхолдера (stakeholder requirement)
b) Нефункциональное требование (non-functional requirement)
c) Бизнес-требование (business requirement)
d) Функциональное требование (functional requirement)

5. Владелец бизнес-процесса — это
a) ответственный исполнитель
b) лицо, которое отвечает за результат процесса, заинтересовано в нем, обладает ресурсами и полномочиями для его выполнения
c) функциональный менеджер
d) спонсор проекта

6. Аналог BPMN-диаграммы в UML — это
a) Диаграмма деятельности (activity diagram)

b) Диаграмма компонентов (Component diagram)
c) Диаграмма классов (Class diagram)
d) Диаграмма состояний (State diagram)

7. Ключевым отличием проекта от процесса является
a) Требования к качеству результата
b) Ограничение в ресурсах
c) Обязательное наличие результата
d) Уникальность

8. Разработка требований к программному продукту в Agile-проектах характеризуется
a) нестабильным характером требований
b) итеративностью циклов детализации требований
c) появлением новых бизнес-потребностей
d) отсутствием ТЗ (технического задания) по ГОСТ

9. Диаграмма Исикавы (рыбья кость) нужна, чтобы
a) показать причинно-следственную связь процессов с результатом
b) определить потенциальные источники проблемы и оценить степень их влияния на результат
c) повысить уровень управляемости бизнес-процессов
d) сформировать полный комплект документации СМК

10. Организационная структура, которая предполагает двойное подчинение, например, начальнику функциональному отдела и менеджеру проекта, называется
a) Функциональная
b) Проектная
c) Процессная
d) Распределенная

Критерии оценки результатов тестирования:

0,5 балл – за каждый правильный ответ.

Максимальное количество баллов за тест – 5 баллов.

Приложение 2

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Комплект тестов
к разделу «Интеллектуальный анализ данных, извлечение знаний из данных»

1. Формат Parquet считается
a) неструктурированным
b) полуструктурированным
c) строковым
d) колоночным (столбцовым)

2. Для машинного обучения подходят данные
a) Любых форматов в цифровом виде
b) Числовые типа int
c) Бинарные
d) Предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов

3. Для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON отлично подходит СУБД
a) HBase
b) Cassandra
c) Hive
d) Elasticsearch

4. Для распределенного глубокого машинного обучения (Deep Learning) больше подходит фреймворк
a) TensorFlow
b) Flask
c) PyTorch
d) Scikit-learn

5. Для реализации микросервисной архитектуры и интеграции разрозненных систем подходит
a) Apache Kafka
b) Apache Spark
c) Apache AirFlow
d) Apache Hadoop

6. Apache NiFi используется для
a) визуализации результатов аналитики
b) эффективного хранения больших данных
c) маршрутизации потоков Big Data и построения ETL-конвейеров
d) оптимизации SQL-запросов к DWH

7. Повысить производительность Apache Kafka можно с помощью:
a) Увеличения плотности разделов на каждом брокере
b) Повышения коэффициента репликации
c) Увеличения размера сообщений
d) Замены HDD-дисков на SSD

8. Автоматизировать запуск пакетных задач в рамках конвейера обработки больших данных по расписанию можно с помощью
a) Apache Hive
b) Apache Hadoop
c) Apache Kafka
d) Apache AirFlow

9. Выберите технологию потоковой обработки событий в режиме реального времени
a) Spark Streaming
b) Apache Kafka
c) Apache Hadoop
d) MapReduce

10. Объём накопленных человечеством цифровых данных на 2012 год измеряется:
a) петабайтами
b) зеттабайтами
c) экзабайтами
d) йоттабайтами

Критерии оценки результатов тестирования:

0,5 балл – за каждый правильный ответ.

Максимальное количество баллов за тест – 5 баллов.

Приложение 3

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Комплект тестов
к разделу «Регрессионный анализ. Кластеризация»

1. Как не используют выборки из генеральной совокупности аналитики больших данных?
а) как метод формирования комплексного суждения о генеральной совокупности случайной величины;
б) как метод тестирования полученных моделей;
в) как метод верификации исходных данных.

2. Укажите лишний этап построения статистической модели:
а) сбор и верификация исходных данных;
б) выбор факторов;
в) построение модели;
г) получение оценок;
д) согласование полученных результатов с заинтересованными лицами;
е) проверка статистической значимости модели.

3. Глубокое обучение включает в себя:
а) регрессионные модели;
б) совокупность различных нейросетевых моделей;
в) методы классификации;
г) градиентный бустинг;
д) обучение с подкреплением.

4. Какой метод верификации исходных данных не применяется для верификации данных о стоимости активов?
а) семантические анализаторы;
б) матрицы граничных значений;
в) конверторы отраслевых классификаторов;
г) наборы решающих правил;
д) проверка данных с использованием колл-центра;
е) тестовые и валидационные выборки.

5. Какие нейронные сети лучше подходят для задач поиска аналога исследуемого объекта?
а) сети Кохонена;
б) сети встречного распространения;
в) RBF-сети на радиальных базисных функциях;
г) любые MLP-нейросети;
д) все вышеперечисленное.

6. Какая проблема решается путем логарифмического шкалирования исходных данных?

а) мультиколлинеарности;
б) робастности;
в) гетероскедастичности;
г) гомоскедастичности.

7. Какие требования к факторам предъявляют классические статистические модели?
а) значимость;
б) независимость;
в) внятная экономическая интерпретация;
г) все вышеперечисленное.

8. Какая технология машинного обучения реагирует на возникновение новых, не описанных ранее ситуаций, получая данные из внешней среды?
а) обучение с подкреплением;
б) обучение с противником;
в) вероятностное прогнозирование;
г) распознавание образов.

9. Как не используют выборки из генеральной совокупности аналитики больших данных?
а) как метод формирования комплексного суждения о генеральной совокупности случайной величины;
б) как метод тестирования полученных моделей;
в) как метод верификации исходных данных

10. В чем состоит стратегия кластеризации?
а) в объединении близких точек многомерного пространства в один объект (кластер) с усредненными характеристиками;
б) разделении множества на части с помощью плоскостей; с) разделении множества на внутренние, или «свои», точки и внешние, или «чужие», точки.

Критерии оценки результатов тестирования:

1 балл – за каждый правильный ответ.

Максимальное количество баллов за тест – 10 баллов.

Приложение 4

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Комплект тестов
к разделу «Визуальный анализ данных»

1. Выберите одно неверное высказывание про MapReduce:
a) интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
b) MapReduce – это две операции: распределения и сборки данных
c) MapReduce был придуман разработчиками Hadoop
d) MapReduce был анонсирован разработчиками Google

2. Какие из следующих технологий СУБД не используют принцип MapReduce
a) Hadoop
b) Cassandra
c) HDInsight
d) Redis

3. Какие СУБД полностью полагаются на оперативную память при хранении информации:
a) Oracle Exalytics
b) SAP HANA
c) BigTable
d) HBase

4. В чём преимущество колоночно-ориентированных СУБД?
a) они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
b) они позволяют динамически дополнять содержание записей новыми полями
c) они имеют более гибкие возможности аналитики
d) они позволяют эффективно делать межколоночные сравнения

5. Для чего аналитику необходима "песочница"?
a) для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций
b) для хранения всех полученных от заказчика данных
c) для построения отчётов о результатах анализа
d) для снижения затрат, связанных с репликацией данных

6. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:
a) Hadoop
b) Data Warehouse
c) "Песочница"

d) Python

7. Выберите верное утверждение:
a) Data Warehouse создаются для проверки гипотез при анализе больших данных
b) "Песочница" используется для снижения нагрузки на основной Data Warehouse
c) каждый Data Warehouse должен содержать "песочницу"
d) "Песочница" необходима для любого процесса аналитики

8. Ниже приведена последовательность этапов проекта аналитики в соответствии с CRISP-DM, укажите первый этап.
a) моделирование (Modeling)
b) внедрение (Deployment)
c) подготовка данных (Data Preparation)
d) понимание бизнеса (Business understanding)

9. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
a) понимание бизнеса (Business understanding)
b) понимание данных (Data Understanding)
c) моделирование (Modeling)
d) оценка (Evaluation)

10. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
a) понимание бизнеса (Business understanding)
b) подготовка данных (Data Preparation)
c) моделирование (Modeling)
d) оценка (Evaluation)

Критерии оценки результатов тестирования:

0,5 балл – за каждый правильный ответ.

Максимальное количество баллов за тест – 5 баллов.

Приложение 5

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Практические задания
к разделу «Введение в интеллектуальный анализ данных»

Задание. Выполните следующие действия:

1. создайте массив №1 на основе данных (выдаются преподавателем) (при создании массива применить функцию «numpy.array», тип данных массива – int);

2. создайте массив №2, размер которого соответствует размеру массива №1 и заполните его интервальными значениями (при создании массива применить функцию «numpy.arange» или функцию «numpy.linspace», диапазоны интервала и шаг (или количество значений) задать самостоятельно, тип данных массива – float);

3. создайте массив №3, размер которого соответствует размеру массива №1, и заполните его случайными элементами, распределёнными по заданному закону (при создании массива применить функцию, включенную в состав пакета «numpy.random» (по вариантам), тип данных массива – float);

4. сделать детальное описание используемой в п.3 функции, т.е. привести теоретическое описание, привести общий вид функции и описать состав её аргументов и возвращаемое значение;

5. осуществите следующие векторные операции (тип данных результирующих массивов – float): - сложить массивы №1, №2, №3 и умножить результат на 5 (скаляр); - вычесть из массива №1 массив №2; - умножить массивы №1, №3; - разделить массив №1 на массив №2; - возвести элементы массива №1 в степень 3.

Задание по вариантам (функции распределения)

№	Функция распределения	№	Функция распределения	№	Функция распределения
1	beta	11	laplace	21	pareto
2	binomial	12	logistic	22	poisson
3	chisquare	13	lognormal	23	power
4	dirichlet	14	logseries	24	rayleigh
5	exponential	15	multinomial	25	standard_cauchy
6	f	16	multivariate_normal	26	standard_exponential
7	gamma	17	negative_binomial	27	standard_gamma
8	geometric	18	noncentral_chisquare	28	standard_normal
9	gumbel	19	noncentral_f	29	standard_t
10	hypergeometric	20	normal	30	triangular

Критерии оценки выполнения заданий:
0 баллов – задание не выполнено;
3 балла – задание выполнен частично, допущен ряд грубых ошибок;
4 балла – задание выполнено полно, допущен ряд неточностей;
5 баллов – задание выполнено правильно и в полном объеме.

Приложение 6

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Практические задания
к разделу «Интеллектуальный анализ данных, извлечение знаний из данных»

Задание

На основе данных об изменении стоимости активов рассчитайте экономико-статистические показатели с целью дальнейшей оценки рисков инвестирования: минимальное значение; максимальное значение; математическое ожидание; дисперсию; стандартное отклонение; среднеквадратическое отклонение и коэффициент вариации. Для выполнения данного задания необходимо выполнить ниже перечисленные действия:

1. создать массив на основании официальных статистических данных о курсе акций за предыдущий месяц (тип данных массива – str, источник данных – https://ru.investing.com/equities/, эмитент акций – по вариантам, применить функции (на выбор): «numpy.loadtxt», «numpy.fromfile»);

2. создать массив, который должен содержать только информацию о ценах за каждый торговый день: ценах открытия, минимальных ценах, максимальных ценах и ценах закрытия (тип данных массива – float, источник данных – созданный в п.1 задания массив, эмитент акций – по вариантам, применить функции (на выбор): «numpy.split», «numpy.split», «numpy.hsplit», использование среза совместно с «numpy.array»);

3. создать массив (вектор-столбец), который должен содержать информацию о средней цене актива за каждый торговый день (применить функции (совместно): «numpy.mean», «numpy.Reshape»);

4. осуществить слияние массивов, созданных на шаге 2 (слева) и шаге 3 (справа) (применить функции (на выбор): «numpy.concatenate», «numpy.hstack»);

5. рассчитать минимальное значение, максимальное значение, математическое ожидание, дисперсию, стандартное отклонение, среднеквадратическое отклонение и коэффициент вариации по всем видам цен активов. Для расчетов показателей использовать следующие функции и способы расчета:

- математическое ожидание («numpy.mean»);
- дисперсия («numpy.mean»);
- стандартное отклонение («numpy.std»);
- среднеквадратическое отклонение (корень квадратный из дисперсии);
- коэффициент вариации (среднеквадратическое отклонение/математическое ожидание));

6. запишите полученный массив на диск (применить функцию: «numpy.save»)

Задание по вариантам

№	Эмитент акций	№	Эмитент акций	№	Эмитент акций
1	ВТБ	11	РУСАЛ	21	Ростелеком
2	ФСК ЕЭС ОАО	12	АК АЛРОСА	22	НОВАТЭК
3	РусГидро	13	Система	23	Детский мир
4	Интер РАО ЕЭС ОАО	14	ММК ОАО	24	ЛУКОЙЛ
5	МКБ	15	НЛМК	25	ТМК
6	Сбербанк	16	Роснефть	26	Северсталь

7	Газпром	17	М.видео	27	ФосАгро
8	Магнит	18	Московская биржа	28	Lenta Ltd
9	Сургутнефтегаз	19	Татнефть	29	X5 Retail Group
10	Аэрофлот	20	МТС	30	Яндекс

Критерии оценки выполнения заданий:
0 баллов – задание не выполнено;
3 балла – задание выполнен частично, допущен ряд грубых ошибок;
4 балла – задание выполнено полно, допущен ряд неточностей;
5 баллов – задание выполнено правильно и в полном объеме.

Приложение 7

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Практические задания
к разделу «Регрессионный анализ. Кластеризация»

Задание

Выполните следующие действия:

1. На основании официальных статистических данных о курсе акции, которые были использованы при решении задания по разделу «Интеллектуальный анализ данных, извлечение знаний из данных» (предварительно необходимо заменить информацию о ценах для нескольких строк на следующие символы «-», « », «н/д»: всего заменить не более 3-4 значений, при этом в одной из строк должно быть не менее 2 таких символов), используя функцию pandas.read_csv, создайте набор данных, в котором решите следующие проблемы:

а) проблему индексирования: определите имена столбцов либо по данным, содержащимся в первой строке исходного файла (аргумент header) или задайте имена столбцов самостоятельно (аргумент names), а также исключите из результирующего набора данных столбец с данными об изменении цен (аргумент usecols);

б) проблему выведения типа и преобразование данных: определите тип данных для значений, содержащихся в столбцах (для столбца с данными о датах установить тип данных datetime64, для цен и объёмов продаж – float64), при необходимости конвертируя значения (аргументы dtype и converters совместно с пунктом б); определите последовательность значений, интерпретируемых как маркеры отсутствующих данных (аргумент na_values);

в) проблему «грязных данных»: определите символ, который нужно распознать как десятичную точку (аргумент decimal).

2. На основании набора данных, полученного в п.1, осуществите:

а) сортировку значений по столбцу с данными об объёмах продаж по убыванию (sort_values ());

б) ранжирование значений по столбцу с данными о цене акции на момент открытия (rank ());

в) получение сводной статистики по динамике цен на акцию (describe());

г) вычисление корреляции и ковариации по цене закрытия и объёму продаж (corr() и cov());

д) обработку отсутствующих данных:

- исключите из набора данных строки, в которых отсутствует 2 и более значений (dropna() с параметром how);

- заполните пустые значения в наборе данных средним арифметическим значением (для ценовых данных – по данным строк, по данным об объёмах продаж – по данным столбца) (dropna() с параметром values);

3. На основании официальных статистических данных о курсе акции, которые были использованы при решении задания по разделу «Интеллектуальный анализ данных, извлечение знаний из данных», а также данных о курсе акции за другой период (например, за прошлый год) осуществите:

а) соединение двух наборов данных в один используя конкатенацию (concat());

б) соединение двух наборов данных в один используя реляционный подход (merge()).

4. На основании официальных статистических данных о курсе акций, которые были использованы при решении задания по разделу «Интеллектуальный анализ данных, извлечение знаний из данных», а также данных о курсе акции за перекрывающий период (например, за текущий год) осуществите:

а) соединение двух наборов данных в один, используя конкатенацию, не допуская дублирования данных (concat());

б) соединение двух наборов данных в один используя реляционный подход, не допуская дублирования данных (merge()).

5. На основании официальных статистических данных о курсе акции, которые были использованы при решении задания по разделу «Интеллектуальный анализ данных, извлечение знаний из данных», а также данных о курсе любой другой акции за аналогичный период осуществите:

а) соединение двух наборов данных в один, используя конкатенацию (concat());

б) соединение двух наборов данных в один используя реляционный подход (merge ()).

Критерии оценки выполнения заданий:

0 баллов – задание не выполнено;

3 балла – задание выполнен частично, допущен ряд грубых ошибок;

4 балла – задание выполнено полно, допущен ряд неточностей;

5 баллов – задание выполнено правильно и в полном объеме.

Приложение 8

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра экономики и организации производства

Практические задания
к разделу «Визуальный анализ данных»

Задание 1.

Выполните следующие действия:

а) создайте Рис. (объект «figure») с двумя областями рисования («объект axes»), расположенными друг под другом;

б) на основании официальных статистических данных о курсе акции, которые были использованы при решении задания по разделу «Интеллектуальный анализ данных, извлечение знаний из данных» в верхней области постройте с помощью линейных графиков («pyplot.plot») ценовой канал по данным ежедневной максимальной и минимальной цены;

в) на основании официальных статистических данных о курсе акции, которые были использованы при решении задания по разделу «Интеллектуальный анализ данных, извлечение знаний из данных»в нижней области отобразите с помощью стержневых графиков («pyplot.stem») динамику ежедневных объёмов торгов.

Задание 2

Выполните следующие действия:

а) создайте Рис. (объект «figure») с двумя областями рисования (объект «axes»), расположенными рядом друг с другом;

б) на основании данных, полученных при решении задания 1 (п.1 и п.2) в левой области постройте диаграмму разброса («pyplot.scatter»); в) на основании данных, полученных при решении задания 1 (п.3) в правой области постройте гистограмму («pyplot.hist»).

Критерии оценки выполнения заданий:

0 баллов – задание не выполнено;

3 балла – задание выполнен частично, допущен ряд грубых ошибок;

4 балла – задание выполнено полно, допущен ряд неточностей;

5 баллов – задание выполнено правильно и в полном объеме.

Приложение 9

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт сервиса и отраслевого управления
Кафедра «Экономики и организации производства»

Вопросы к зачету по дисциплине
«Технологии интеллектуального анализа BigData в экономических исследованиях»

1. Какие тренды информационно-коммуникационных технологий способствовали развитию Data Mining?
2. Приведите примеры применения методов Data Mining для решения практических задач.
3. Какие области человеческой деятельности наиболее и наименее подходят для их анализа методами Data Mining?
4. Что понимается под Data Mining и Big Data? Почему возникла такая терминология?
5. В чем состоит суть индуктивных и дедуктивных подходов в Data Mining?
6. Каковы основные этапы интеллектуального анализа данных?
7. Какие классификации методов Data Mining существуют? Приведите примеры.
8. В чем заключается предварительная обработка данных и какова ее цель? Какие подходы при этом применяются?
9. В чем заключается оптимизация признакового пространства? Какие методы с трансформацией и без трансформации пространства применяют и в чем их отличия?
10. В чем заключается метод классификации? Какие подходы для его реализации могут быть использованы и в чем их суть?
11. Что такое неконтролируемая классификация и какие методы применяют для ее реализации?
12. В чем заключается суть метода машины опорных векторов и в чем его преимущество перед аналогами?
13. Как работают деревья принятия решений? Какие их разновидности существуют? Каковы пределы применимости этого метода?
14. Что такое регрессия? Какие подходы применяют для ее реализации?
15. Как работают ассоциативные алгоритмы?
16. Как работают алгоритмы последовательной ассоциации?
17. Что такое обнаружение аномалий? Приведите примеры применения этого подхода и методы его реализации.
18. Что такое визуализация и какие инструменты ее реализации существуют?
19. Какие инструменты, модели и технологии существуют сегодня для реализации высокопроизводительных вычислений? Какие критерии эффективности при этом используют?
20. Приведите примеры коммерческих многофункциональных систем и свободно распространяемых решений, реализующих инструментарий Data Mining. Их сравнительные характеристики.

Критерии оценки:

- балл 61-100 (зачтено) выставляется обучающемуся, если он показал всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;

- балл 0-60 (не зачтено) выставляется обучающемуся, если он не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.