МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
**Федеральное государственное бюджетное образовательное учреждение высшего образования**
**К Г Э У** **«КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ ЭНЕРГЕТИЧЕСКИЙ УНИВЕРСИТЕТ»**
(ФГБОУ ВО «КГЭУ»)

# ОЦЕНОЧНЫЕ МАТЕРИАЛЫ
## для проведения текущего контроля успеваемости и промежуточной аттестации студентов по итогам освоения дисциплины

## Инжиниринг данных

Направление подготовки 09.04.01 Информатика и вычислительная техника

Направленность (профиль) Инженерия искусственного интеллекта

Квалификация Магистр

Форма обучения Очная

Составлено автором:

| № п/п | Фамилия Имя Отчество | Ученая степень, ученое звание | Должность | Подразделение |
|---|---|---|---|---|
| 1 | Созыкин Андрей Влаимирович | кандидат технических наук | доцент | Кафедра информационных технологий и систем управления, ИРИТ-РТФ, УрФУ |

Оценочные материалы оформлены в соответствие с ПОЛОЖЕНИЕМ О ПОРЯДКЕ РАЗРАБОТКИ И УТВЕРЖДЕНИЯ ОБРАЗОВАТЕЛЬНЫХ ПРОГРАММ – ПРОГРАММ БАКАЛАВРИАТА, ПРОГРАММ СПЕЦИАЛИТЕТА И ПРОГРАММ МАГИСТРАТУРЫ В КГЭУ

## 1. Цель и задачи текущего контроля и промежуточной аттестации студентов по дисциплине «Инжиниринг данных»

*Цель текущего контроля* - систематическая проверка степени освоения программы дисциплины «Инжиниринг данных», уровня сформированности знаний, умений, навыков, компетенций на текущих занятиях

*Задачи текущего контроля:*

1. определение индивидуального учебного рейтинга студентов;
2. своевременное выполнение корректирующих действий по содержанию и организации процесса обучения; обнаружение и устранение пробелов в усвоении учебной дисциплины;
3. подготовки к промежуточной аттестации.

В течение семестра при изучении дисциплины реализуется комплексная система поэтапного оценивания уровня освоения – балльно-рейтинговая система. За каждый вид учебных действий студенты получают определенное количество баллов. В течение семестра студент может набрать до 60-ти баллов.

*Цель промежуточной аттестации* - проверка степени усвоения студентами учебного материала за время изучения дисциплины, уровня сформированности компетенций после завершения изучения дисциплины. Аттестация проходит в форме зачета.

*Задачи промежуточной аттестации:*

1. определение уровня усвоения учебной дисциплины;
2. определение уровня сформированности компетенций.

## 2. Основное содержание текущего контроля и промежуточной аттестации студентов

В результате изучения дисциплины «Инжиниринг данных» формируются следующие компетенции или их составляющие:

### 2.1. Основное содержание текущего контроля

| Коды компетенций | Совокупность ожидаемых результатов образования студентов в форме компетенций по завершении модуля / освоения дисциплины | Содержание оценочных заданий для выявления сформированности компетенций у студентов по завершении модуля/освоения дисциплины | | |
|---|---|---|---|---|
| | | Базовый уровень | Продвинутый уровень | Высокий уровень |
| 1 | 2 | 3 | 4 | 5 |
| ОПК-3 | ОПК-3.1. Анализирует профессиональную информацию, обосновывает выводы и рекомендации по решению профессиональных задач<br>ОПК-3.2. Составляет научные доклады, публикаций, аналитические обзоры в сфере профессиональной деятельности | Решение практического задания | Решение практического задания, написание отчета | Решение практического задания, написание отчета, ответы на вопросы |

### 2.2. Основное содержание промежуточной аттестации студентов

| Коды ком-петенций | Совокупность ожидаемых результатов образования студентов в форме компетенций по завершении модуля / освоения дисциплины | Содержание оценочных заданий для выявления сформированности компетенций у студентов по завершении модуля/освоения дисциплины | | |
|---|---|---|---|---|
| | | Базовый уровень | Продвинутый уровень | Высокий уровень |
| 1 | 2 | 3 | 4 | 5 |
| ОПК-3 | ОПК-3.1. Анализирует профессиональную информацию, обосновывает выводы и рекомендации по решению профессиональных задач<br>ОПК-3.2. Составляет научные доклады, публикаций, аналитические обзоры в сфере профессиональной деятельности | Решение кейса | Решение кейса | Решение кейса |

## 3. Оценочные средства для текущего контроля успеваемости и промежуточной аттестации по итогам освоения дисциплины

| п/п | Наименование компоненты | Критерии оценки | Максимальное число баллов |
|---|---|---|---|
| 1 | Практические занятия - 4 занятия | выполнение задания 8 баллов<br>оформление выводов 4 балла<br>ответы на вопросы 3 балла | 15 баллов за каждое занятие<br>**60 баллов** |
| 2 | Решение кейса | определен способ преобразования данных в соответствие с решаемой задачей 5 баллов<br>исходные данные преобразованы в набор данных, пригодных для анализа данных 10 баллов<br>проведена разметка и очистка данных 10 баллов<br>проведена подготовка данных 15 баллов | **40 баллов** |
| | | ИТОГО | **100 баллов** |

***Содержание практических заданий***

**Практическое задание 1. Библиотеки для работы с данными**

Написать код, снабжая его комментариями.
1. Открыть набор данных о футболистах data_sf.csv с помощью функции read_csv модуля pandas и применить к нему метод .info(). Ответить на вопросы:

1. Сколько всего колонок в наборе данных о футболистах?
2. Сколько непустых элементов в колонке Club?
3. Какой тип данных у колонки Wage?
4. Какой тип данных у колонки Position?
5. Сколько колонок имеют тип object?
6. Какая колонка будет последней в наборе данных?

7. Есть ли в наборе данных колонка с названием 'Surname'?

2. Конвертировать набор данных об абитуриентах br11er25_ab.json. Ответить на вопросы:
   1. Сколько всего колонок в наборе данных?
   3. Какой тип данных у колонки Numbre?
   4. Какой тип данных у колонки Home?
   5. Сколько колонок имеют типы object?
   6. Какая колонка будет последней в наборе данных?

3. Открыть файл о калорийности продуктов kaloriy.html. Ответить на вопросы:
   1. Сколько всего колонок в наборе данных?
   3. Какой тип данных у колонки Product?
   4. Какой тип данных у колонки Day?
   5. Сколько колонок имеют типы object?
   6. Какая колонка будет последней в наборе данных?

**Практическое задание 2. Работа с базами данных в Python. Работа с изображениями, видео и звуковыми файлами**

Написать код, снабжая его комментариями.

1. Подключиться к БД dbRtf. Выполнить запрос к таблице Price и получить все данные

Ответить на вопросы:
   1. Сколько всего колонок в наборе данных?
   2. Сколько непустых элементов в колонке Old?
   3. Какой тип данных у колонки Data?
   4. Какое последнее значение в третьей колонке?

2. Открыть файл, содержащий наборы изображений. Изображения имеют размерность (3000, 3000, 1). Значение пикселей лежат в диапазоне 0...255. Просмотреть первые 10 изображений. Преобразовать в набор данных

3. Открыть файл, содержащий наборы изображений. Изображения имеют размерность (300, 300, 3). Значение пикселей лежат в диапазоне 0...255. Просмотреть первые 10 изображений. Преобразовать в набор данных

4. Повторить аналогичные действия с получением набора данных из звуковых файлов.

**Практическое задание 3. Сбор данных и формирование набора данных**

Написать код, снабжая его комментариями.

1. Открыть набор данных о футболистах data_sf.csv с помощью функции read_csv модуля pandas.

2. Для ответа на дальнейшие задачи нужно применить к набору данных о футболистах метод .describe() (возможно, с какими-то параметрами – параметры изучить самостоятельно).
   1. Каково среднее значение возраста футболистов в наборе данных?
   2. Каков минимальный возраст футболиста?
   3. Какова максимальная заработная плата за год (Wage) у футболистов?
   4. Какова медианная заработная плата за год (Wage) у футболистов?
   5. Каково минимальное значение параметра 'Penalties'?
   6. Какое значение у первого (25%) квартиля параметра ShortPassing?
   7. Какова самая частая национальность (Nationality) футболистов?
   8. Сколько разных клубов в наборе данных о футболистах?
   9. Сколько раз встречается самая частая позиция 'GK' в наборе данных?

10. Каков максимальный возраст футболиста?

3. Для ответа на дальнейшие задачи нужно применить статистические методы
   1. Чему равен средний возраст (Age), футболистов в наборе данных, округлённый до целого?
   2. Каково количество непустых строк в колонке Composure (Хладнокровие) набора данных о футболистах?
   3. Каково в наборе данных о футболистах стандартное отклонение параметра коротких пасов (ShortPassing), округлённое до второго знака после запятой?
   4. Какова сумма заработных плат за год (Wage) в наборе данных о футболистах?
   5. Какова минимальная стоимость футболиста (Value) в наборе данных о футболистах?
4. Для ответа на дальнейшие задачи нужно применить методы фильтрации
   1. Какова средняя скорость (SprintSpeed) футболистов, зарплата (Wage) которых выше среднего?
   2. Какова средняя скорость (SprintSpeed) футболистов, зарплата (Wage) которых ниже среднего?
   3. Какую позицию (Position) занимает футболист с самой высокой зарплатой (Wage)?
   4. Сколько пенальти (Penalties) забили бразильские (Nationality, Brazil) футболисты за период, данные о котором представлены в датасете?
   5. Укажите средний возраст (Age) игроков, у которых точность удара головой (HeadingAccuracy) > 50.
   6. Укажите возраст (Age) самого молодого игрока, у которого хладнокровие (Composure) и реакция (Reactions) превышают 90% от максимального значения, представленного в датасете.
   7. Определите, насколько средняя реакция (Reactions) самых взрослых игроков (т.е. игроков, чей возраст (Age) равен максимальному) больше средней реакции самых молодых игроков.
   8. Из какой страны (Nationality) происходит больше всего игроков, чья стоимость (Value) превышает среднее значение?
   9. Определите, во сколько раз средняя зарплата (Wage) голкипера (Position, GK) с максимальным значением показателя " Рефлексы" (GKReflexes) выше средней зарплаты голкипера с максимальным значением показателя "Владение мячом" (GKHandling).
   10. Определите, во сколько раз средняя сила удара (ShotPower) самых агрессивных игроков (игроков с максимальным значением показателя "Агрессивность" (Aggression)) выше средней силы удара игроков с минимальной агрессией. Ответ округлите до сотых.

**Практическое задание 4. Методы очистки и подготовки данных. Очистка и подготовка данных на Python. Разметка данных.**

Написать код, снабжая его комментариями.

1. Открыть файл log.csv. Ознакомиться с данными.
2. Посмотрите все уникальные значения в sample.csv
3. Посмотрите, сколько непустых значений в колонке City
4. Файл log.csv, в колонке user_id есть записи, которые содержат технические ошибки. Укажите, что записано в поле user_id в строчках с ошибкой
5. Создайте новый датафрейм sample2, в который будут входить только записи о людях в возрасте меньше 30 лет

6. Создайте новый датафрейм log_win, в который будут входить только записи, где пользователь выиграл. Посчитайте, сколько таких записей, и сохраните в переменной win_count
7. Создайте новый датафрейм sample2, в который будут входить только записи о рабочих младше 30 лет
8. Найдите записи, где в городах нет буквы "o", и сохраните в переменную sample4. Не забудьте про NaN и параметр na
9. Сохраните в переменную new_log датафрейм, из которого удалены записи с ошибкой в поле user_id
10. С помощью apply и лямбда-функции увеличьте возраст во всех записях на 1 год и сохраните в sample2. В переменной sample2 должен содержаться весь датафрейм sample
11. С помощью apply и lambda-функции замените все буквы в поле City на маленькие и сохраните в sample2. Вам может понадобиться функция s.lower()
12. Напишите функцию profession_code, которая на вход получает строку, а на выход возвращает:
- 0 — если на вход поступила строка "Рабочий"
- 1 — если на вход поступила строка "Менеджер"
- 2 — в любом другом случае

13. Напишите функцию age_category, которая на вход получает число, а на выход отдаёт:
- "молодой" — если возраст меньше 23
- "средний" — если возраст от 23 до 35
- "зрелый" — если возраст больше 35

***Вопросы для подготовки к зачету по курсу «Инжиниринг данных»***

1. Библиотека pandas в Python.
2. Работа с данными в формате CSV в Python.
3. Работа с данными в формате JSON в Python.
4. Работа с данными в формате HTML в Python.
5. Работа с изображениями в Python.
6. Работа с видео в Python.
7. Работа с аудио в Python.
8. Работа с Parquet в Python.
9. Работа с графами знаний в Python.
10. Этапы и инструменты создания наборов данных для машинного обучения.
11. Загрузка данных с Web-сайтов.
12. Загрузка данных из социальных сетей.
13. Методы и инструменты подготовки данных.
14. Методы и инструменты очистки данных.
15. Разметка данных.
16. Общедоступные платформы для хранения данных.
17. Архитектура центров обработки данных.
18. Кластеры для параллельных и распределенных вычислений.
19. Экосистема для распределенного хранения и обработки больших объемов данных: Apache Hadoop.
20. Распределенная файловая система HDFS.
21. Распределенная обработка данных в Apache Spark.
22. Работа с данными с использованием Apache Spark DataFrame.
23. Источники данных для Apache Spark DataFrame.
24. Обработка данных в Apache Spark DataFrame.
25. Использование SQL в Apache Spark DataFrame.

***Пример типового кейса для проведения промежуточной аттестации***

1. Исходные данные представлены в файле Excel. Открыть файл и просмотреть структуру данных.

2. Сформировать набор данных. Набор данных должен быть представлен в виде датафрейма. Имена столбцов не должны содержать пробелов. Типы данных должны соответствовать значениям.

3. Проверить данные на наличие ошибок и пустых значений. Просмотреть ошибки. Преобразовать, либо удалить данные в зависимости от ошибок и пропусков. Подумайте, что следует сделать, если данные чрезвычайно разреженные по определенным признакам.

4. Выявить выбросы и удалить соответствующие записи.

5. Построить два частотных распределения, два графика и две диаграммы box-whiskies. По каждому рисунку сделать выводы.