МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ УНИВЕРСИТЕТ

ВЫСШАЯ ШКОЛА ЦИФРОВЫХ ТЕХНОЛОГИЙ

Кафедра «Интеллектуальных систем и технологий»

Экзаменационная работа

по дисциплине «Большие данные»

Вариант 5

**Выполнил:**

студент группы ИСТ-21-1

Бокий М.М.

**Проверил:**

Ассистент

Горбунова А.Д.

Тюмень, 2024

Задание: загрузить датасет по ссылке <https://disk.yandex.ru/d/okMgNUtLmVDYBQ>. С помощью Apache Spark загрузить датасет, проанализировать признаки, выполнить необходимые этапы предобработки данных. Решить задачу прогнозирования психологического состояния человека (столбец «Psychological State»). Подобрать модель классификации и её гиперпараметры.

Листинг exam.py

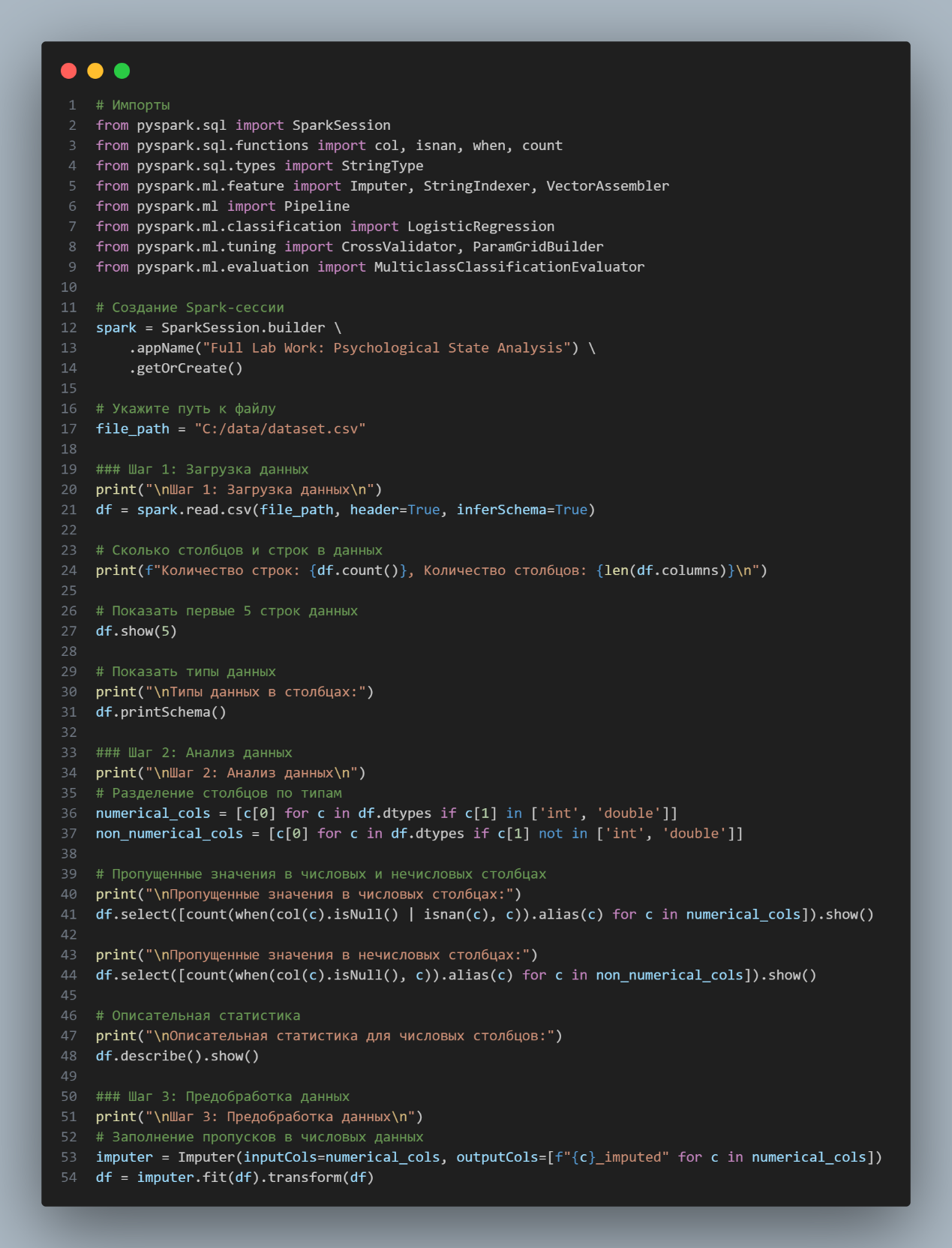


Рисунок 1. Листинг ПО exam.py



Рисунок 2. Листинг ПО exam.py

Вывод файла exam.py

**Шаг 1: Загрузка данных**

1. **Что делается?**

* Загружается датасет из файла (в данном случае dataset.csv).
* Проверяется, сколько строк и столбцов в датасете.
* Выводятся первые несколько строк данных и типы столбцов (например, числа, текст или дата).

1. **Для чего это нужно?**

* Понять, какие данные у нас есть.
* Убедиться, что данные загружены правильно, и получить представление о структуре таблицы.

Изображение выглядит как текст, снимок экрана, черный, Шрифт

Автоматически созданное описание

Рисунок 3. Вывод шага 1-го.

Изображение выглядит как текст, снимок экрана, Шрифт, черно-белый

Автоматически созданное описание

Рисунок 4. Вывод шага 1-го, отображение типа данных

**Шаг 2: Анализ данных**

1. **Что делается?**

* Определяется, какие столбцы числовые, а какие текстовые.
* Проверяются пропущенные значения: сколько данных отсутствует в каждом столбце.
* Выводится описательная статистика для числовых столбцов (например, минимальное, максимальное, среднее значение).

1. **Для чего это нужно?**

* Анализ помогает выявить проблемы с данными, например, пропуски или аномалии.
* Статистика показывает, насколько разнообразны данные и есть ли значения, которые нужно обработать (например, выбросы).

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Рисунок 5. Отображение вывода 2-го шага, пропущенные значения

Изображение выглядит как текст, снимок экрана, черный

Автоматически созданное описание

Рисунок 6. Отображение вывода 2-го шага, описательная статистика для числовых столбцов

**Шаг 3: Предобработка данных**

1. **Что делается?**
   * Пропущенные значения в числовых столбцах заполняются средними значениями, чтобы модель могла использовать эти данные.
   * Текстовые данные преобразуются в числа, потому что модели машинного обучения работают только с числами.
   * Все признаки (X) собираются в единый вектор features, а целевая переменная (Psychological State) используется как метка (Y).
2. **Для чего это нужно?**
   * Без обработки данные нельзя использовать в модели.
   * Преобразование данных делает их пригодными для машинного обучения.

Изображение выглядит как снимок экрана, текст, черно-белый, черный

Автоматически созданное описание

Рисунок 7. Отображение вывода 3-го шага, предобработка данных

**Шаг 4: Обучение модели**

1. **Что делается?**

Данные делятся на две части:

* + - **Обучающая выборка:** для того, чтобы модель "училась" на данных.
    - **Тестовая выборка:** для проверки точности модели.
  + Обучается модель логистической регрессии, которая пытается предсказать целевую переменную Psychological State на основе признаков (X).
  + Модель делает прогнозы на тестовых данных.

1. **Для чего это нужно?**
   * На этом этапе создаётся модель, которая учится находить закономерности в данных.
   * Прогнозы показывают, насколько хорошо модель справляется с задачей.

Изображение выглядит как текст, снимок экрана, черно-белый, черный

Автоматически созданное описание

Рисунок 8. Отображение вывода 4-го шага, обучение модели

**Шаг 5: Подбор гиперпараметров**

1. **Что делается?**

Подбираются параметры модели, чтобы улучшить её точность. Например:

* + - **Регуляризация (regParam)**: помогает предотвратить переобучение.
    - **Эластичная сеть (elasticNetParam)**: регулирует баланс между двумя методами регуляризации.
  + Используется кросс-валидация, чтобы проверить, какие параметры дают лучшие результаты.
  + После подбора модель делается более точной.

1. **Для чего это нужно?**
   * Без настройки параметров модель может работать неэффективно.
   * Кросс-валидация позволяет улучшить модель и добиться максимально возможной точности.

Изображение выглядит как текст, снимок экрана, Шрифт, дизайн

Автоматически созданное описание

Рисунок 9. Отображение вывода 4-го шага, обучение модели

Ответы на вопросы:  
**1. Что такое данные? Что такое большие данные? Основные характеристики больших данных. Сравнительная оценка традиционных баз данных и больших данных. Основные проблемы при обработке больших данных.**

* **Данные** — это любая информация, которая может быть записана, сохранена и использована.
* **Большие данные** — это огромные объемы информации, которые сложно обработать стандартными инструментами. Примеры: данные из соцсетей, видео, GPS-локации.
* **Характеристики больших данных**: объем (Volume), скорость (Velocity), разнообразие (Variety), достоверность (Veracity), ценность (Value).
* **Традиционные базы данных**: работают с структурированной информацией, медленны для обработки больших объемов.
* **Большие данные**: могут работать с неструктурированной информацией, требуют специальных инструментов (например, Hadoop, Spark).
* **Проблемы**: скорость обработки, качество данных, хранение, интеграция.

**2. Что такое модель? Основные задачи, решаемые в процессе моделирования. Технологии, применяемые при обработке больших данных.**

* **Модель** — это упрощенное представление реального мира (например, математическое, статистическое).
* **Задачи моделирования**: прогнозирование, анализ, автоматизация процессов.
* **Технологии для обработки больших данных**: Hadoop, Spark, базы данных (NoSQL, SQL), инструменты машинного обучения.

**3. Что такое pipeline и workflow? Их основные отличия. Виды pipeline в обработке данных и машинном обучении. Построить pipeline обработки данных.**

* **Pipeline** — это последовательность действий для обработки данных (например, загрузка → очистка → анализ).
* **Workflow** — это более сложный процесс, включает взаимодействие между разными задачами.
* **Различия**: pipeline — линейный процесс, workflow — более сложный и разветвленный.
* **Виды pipeline**: для ETL (Extract, Transform, Load), обработки данных, обучения моделей.
* **Пример pipeline**:
  1. Загрузить данные.
  2. Очистить данные.
  3. Обучить модель.
  4. Сохранить результаты.

**4. Автоматизация процесса сбора, подготовки и загрузки данных. Инструменты оркестрации. Apache Airflow, как инструмент оркестрации. Как устанавливается? Как выглядит интерфейс? Основы построения DAGs.**

* **Автоматизация** позволяет сократить время на повторяющиеся задачи.
* **Инструменты**: Apache Airflow, Prefect, Luigi.
* **Apache Airflow**:
  + Используется для управления задачами (сбор, загрузка, обработка).
  + Установка через pip.
  + Интерфейс: веб-приложение с расписанием задач.
  + **DAG** (граф задач): набор связанных действий.

**5. Технологии хранения больших данных: базы данных, хранилища данных и озера данных. Их сравнение и взаимосвязь. Реляционные базы данных и их примеры. Нереляционные базы данных: типы, особенности организации хранения данных и примеры.**

* **Базы данных**: для структурированных данных (MySQL, PostgreSQL).
* **Хранилища данных**: для анализа данных (Snowflake, Google BigQuery).
* **Озера данных**: для хранения необработанных данных (Hadoop, Amazon S3).
* **Реляционные базы данных**: структурированные таблицы, SQL (Oracle, MySQL).
* **Нереляционные базы данных**: NoSQL, для больших объемов и гибкости (MongoDB, Redis).

**6. Технология MapReduce. Применение её в средствах хранения и обработки больших данных. Программные каркасы и библиотеки проекта Hadoop.**

* **MapReduce**: разбивает задачи на две части — Map (распределение) и Reduce (суммирование).
* **Применение**: обработка больших объемов данных.
* **Hadoop**: программный каркас, включает HDFS (хранилище) и MapReduce.

**7. Средства массово-параллельной обработки данных. Spark: особенности его архитектуры, принципы организации параллельной обработки данных, особенности построения сессий.**

* **Spark**: система обработки данных.
* **Особенности**: быстрее, чем Hadoop, работает в памяти.
* **Параллельная обработка**: данные разбиваются на блоки и обрабатываются одновременно.
* **Сессия**: точка входа для работы со Spark.

**8–12. Spark и его технологии обработки данных (тексты, векторы, категории, масштабирование).**

* **Тексты**: обработка через Spark NLP.
* **Векторы**: работа через библиотеки MLlib (например, Word2Vec).
* **Категории**: преобразование категорий в числа (OneHotEncoder).
* **Масштабирование**: нормализация данных для машинного обучения.

**13. Развертывание моделей машинного обучения при создании веб-приложений: библиотеки, которые могут использоваться, их особенности, этапы встраивания моделей машинного обучения в веб-приложение.**

* **Библиотеки**: Flask, Django, FastAPI.
* **Этапы**:
  1. Обучить модель.
  2. Сохранить её.
  3. Встроить через API в веб-приложение.

**14. Контроль версионности моделей. Что такое версии моделей? Зачем их контролировать? Инструменты для контроля версионности моделей. Описание этапов изменения версии моделей машинного обучения.**

* **Версии моделей**: это разные варианты одной модели.
* **Зачем?**: для улучшений и откатов к старым версиям.
* **Инструменты**: MLflow, DVC.
* **Этапы**:
  1. Создать новую версию.
  2. Протестировать.
  3. Обновить или вернуть предыдущую.