Наша цель - разработать многослойную нейронную сеть для классификации методик лечения рака легких на основе медицинских обследований. Модель будет предсказывать класс методики лечения и уровень уверенности врача в диагнозе и выбранной методике лечения.

**Шаги Разработки Модели:**

1. **Анализ и Подготовка Данных:**
   * Изучить предоставленный датасет, определить полезные признаки и подготовить данные для обучения модели.
   * Преобразовать категориальные признаки в числовой формат.
   * Разделить данные на обучающую, валидационную и тестовую выборки.
2. **Разработка Модели:**
   * Создать архитектуру нейронной сети, используя слои, подходящие для классификации.
   * Определить функции потерь и метрики, соответствующие задачам классификации и предсказания уверенности.
3. **Обучение и Валидация:**
   * Обучить модель на обучающей выборке.
   * Настроить гиперпараметры и провести валидацию модели.
4. **Оценка Производительности:**
   * Проверить эффективность модели на тестовой выборке.
   * Использовать соответствующие метрики для оценки классификации и уровня уверенности.
5. **Интеграция с Интерфейсом Бота:**
   * Подготовить функции для приема вводимых данных и получения результатов от модели.
   * Интегрировать модель в телеграм-бота.

**Датасет**: содержит различные признаки, которые можно использовать для обучения нейронной сети. Вот основные столбцы и их характеристики:

1. **id\_ответа**: Уникальный идентификатор ответа.
2. **Раса, Пол, Возраст, Статус курения, ECOG**: Различные демографические и клинические характеристики пациентов.
3. **Есть опухолевая нагрузка? (симптомная опухоль), Ко-мутации (KRAS, p53, STK11 и др.)**: Медицинские характеристики состояния здоровья пациента.
4. **Срок от окончания ХЛТ, Молекулярный статус, PD-L1 статус**: Дополнительные медицинские данные.
5. **Предпочтение пациента по ответу на терапию**: Предпочтения пациента относительно лечения.
6. **Ответ эксперта (Лактионов)**: Рекомендация лечения, это будет нашей целевой переменной для классификации.
7. **Уверенность доктора (Поставьте галочку, если уверены на 100%, Если 75%, Если 50%)**: Уровни уверенности врача в диагнозе и выбранной методике лечения.

**Подход к Созданию Модели:**

1. **Предобработка Данных:**
   * Кодирование категориальных переменных.
   * Нормализация числовых данных, если таковые имеются.
   * Обработка пропущенных значений, если они есть.
2. **Разделение Данных:**
   * Разделение на обучающую, валидационную и тестовую выборки.
3. **Построение Модели:**
   * Использование многослойной нейронной сети (например, с использованием Keras).
   * Входной слой должен соответствовать количеству признаков.
   * Выходной слой для классификации методик лечения.
   * Дополнительный выход для уверенности доктора.
4. **Компиляция и Обучение Модели:**
   * Выбор функции потерь (категориальная перекрестная энтропия для классификации, MSE для уверенности).
   * Выбор оптимизатора (например, Adam).
   * Обучение модели на обучающей выборке с валидацией на валидационной выборке.
5. **Оценка и Тестирование Модели:**
   * Оценка модели на тестовой выборке.

приступим к базовой предобработке данных и созданию нейронной сети для классификации методик лечения рака легких. Этапы предобработки данных .

1. **Кодирование Категориальных Переменных:** Преобразование категориальных признаков в числовые значения, чтобы их можно было использовать в нейронной сети.
2. **Нормализация Данных:** Масштабирование числовых признаков для улучшения процесса обучения.
3. **Обработка Пропущенных Значений:** Заполнение или удаление строк с пропущенными значениями.
4. **Разделение На Обучающую, Валидационную и Тестовую Выборки:** Разделение датасета на части для обучения, валидации и тестирования модели.

Давайте начнем с кодирования категориальных переменных и нормализации данных. Затем мы создадим и обучим модель, используя библиотеку Keras. После этого мы оценим её производительность на тестовой выборке.

Прежде всего, выполним предобработку данных.

Предобработка данных была успешно выполнена. Мы преобразовали категориальные признаки в числовые с помощью one-hot кодирования и нормализовали числовые признаки. Датасет был разделен на обучающую и тестовую выборки. Размеры выборок следующие:

* Обучающая выборка: 199065 записей
* Тестовая выборка: 49767 записей
* Количество признаков после предобработки: 36

Чтобы работать с телеграмм ботом нам будет необходимо сохранить и скачать обученную модель в формате .h5 из Google Colab на ваш локальный компьютер.

**Сохранение Модели:**.

model.save('my\_model.h5')

Этот код сохранит модель в файл с именем my\_model.h5 в текущей рабочей директории Colab.

 **Скачивание Файла:**

python

from google.colab import files

files.download('my\_model.h5')

Скачает файл на локальный комп

Поскольку задача представляет собой многоклассовую классификацию, мы будем использовать softmax в выходном слое, чтобы получить распределение вероятностей по классам. Для обучения модели мы используем функцию потерь categorical cross-entropy.