**Задача 2.**

Используя данные датасета, размещенного по адресу https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset, реализовать задачу прогнозирования сахарного диабета на основе вводимых данных.

Необходимо подготовить данные используя функции библиотеки pandas - dropna() или fillna(),перевести все тестовые данные в единый регистр для однообразия данных и числовой формат, убрать выбросы, используя межквартильный размах или Z-оценку, убрать дубликаты. Поделить на тестовую и обучающие выборки.

Проведите масштабирование данных.

Выбрать один из методов - логистическая регрессия или дерево решений.

Оценить качество полученной модели:

accuracy\_score - оценка точности классификации. Рассчитывается для того, чтобы ограничить количество ложных срабатываний;

f1\_score - среднее значение Точности измерений (Accuracy) и Отзыва (Recall) с Весами (Weight) при наличии.

Оформить выводы.

В процессе работы свои выводы обосновываете графиками.

Теоретические материалы:

Логистическая регрессия — это метод анализа данных, который использует математику для поиска взаимосвязей между двумя факторами данных. Затем эта взаимосвязь используется для прогнозирования значения одного из этих факторов на основе другого. Прогнозирование обычно имеет конечное количество результатов, например «да» или «нет». Это один из нескольких различных методов регрессионного анализа, которые специалисты по обработке данных обычно используют в машинном обучении.

В целом логистическая регрессия исследует, как переменные влияют на одну зависимую переменную, исторические значения обеих переменных.

Логитная функция, независимо от значений независимой переменной, возвращает только значения от 0 до 1 для зависимой переменной. Так логистическая регрессия оценивает значение зависимой переменной.

Существует три подхода к логистическому регрессионному анализу, основанному на результатах зависимой переменной:

− Бинарная логистическая регрессия;

− Многочленная логистическая регрессия;

− Порядковая логистическая регрессия.

Бинарная логистическая регрессия хорошо подходит для задач бинарной классификации, которые имеют только два возможных результата. Зависимая переменная может иметь только два значения, например «yes» и «no» или 0 и 1 Несмотря на то, что логистическая функция вычисляет диапазон значений от 0 до 1, модель бинарной регрессии округляет ответ до ближайших значений. Обычно ответы ниже 0,5 округляются до 0, а ответы выше 0,5 – до 1, так что логистическая функция возвращает двоичный результат.

Многочленная регрессия может анализировать проблемы, которые имеют несколько возможных результатов, если количество результатов ограничено. Многочленная логистическая регрессия работает путем сопоставления значений результатов с различными значениями от 0 до 1 Поскольку логистическая функция может возвращать диапазон непрерывных данных, таких как 0,1, 0,11, 0,12 и т. д., многочленная регрессия также группирует выходные данные с максимально близкими возможными значениями.

Порядковая логистическая регрессия, или упорядоченная логитная модель, является особым типом многочленной регрессии для задач, в которых числа представляют собой ранги, а не фактические значения.

Точность результатов расчёта логистической регрессии целиком и полностью зависит от выборки, на основании которой рассчитывались коэффициенты в уравнении логистической регрессии. Таким образом, построенная модель требует проверки её адекватности. Простейшим способом оценки адекватности модели является проверка этой модели на исходных данных и сравнение полученных данных с предварительно используемыми исходами («исходными исходами»).

Оценка выражается как процент наблюдений с исходами, верно предсказанными с помощью модели регрессии.

Проверка значимости отличия коэффициентов от нуля проводится при помощи статистики Вальда, использующей распределение хи-квадрат и представляющей собой квадрат отношения соответствующего коэффициента к его стандартной ошибке.

Последовательный статистический анализ А. Вальда

С помощью последовательного статистического анализа А. Вальда прогнозировались результаты лечения пациентов, у которых первыми обнаружили туберкулез, а также риск развития заболевании. Данный метод хорошо работает в случае отсутствия части признаков.

Особенность метода заключается в том, что исследование проводят последовательно и после каждого наблюдения делают расчеты, на основании которых приходят к необходимым выводам или устанавливают необходимость продолжить наблюдение. Исследование прекращают выясняется, что полученные данные обеспечивают надежную диагностику.

Благодаря этому число наблюдений, необходимых для обоснованного вывода, значительно сокращается по сравнению с методами, требующими заранее установленного числа наблюдений.

Дерево решений

Деревья принятия решений (DTS) – это непараметрический контролируемый метод обучения, используемый для классификации и регрессии.

Цель метода в создании модели, которая предсказывает значение целевой переменной путем изучения простых правил принятия решений, выводимых из характеристик данных. Дерево можно рассматривать как кусочно-постоянное приближение.

Алгоритм дерева решений основан на условных вероятностях и дает возможность создавать точные и интерпретируемые модели с относительно небольшим вмешательством пользователя. Также деревья решений генерируют правила. Правилом называется условная инструкция, которая может быть понятна людям и использоваться в базе данных для идентификации набора записей. Они обеспечивают прозрачность модели и используются для проверки моделей в таких задачах.

Библиотеки Python:

NumPy — библиотека для работы с многомерными массивами и матрицами;

Pandas — библиотека для работы с табличными данными, включая чтение и запись данных из различных форматов;

Matplotlib — библиотека для визуализации данных, включая построение графиков, диаграмм и др.;

Datetime — библиотека, которая предоставляет удобные инструменты для работы с датами и временем;

Seaborn — библиотека, которая предоставляет мощные инструменты для визуализации данных;