Расчетно-аналитическое задание №1

Цель: применить на практике методы бинарной классифкации.

Задачи: Построить модель логистической регрессии, Оценить их качество.

Этапы

1. Выполните импорт библиотек
2. Загрузить данные из файла diabets
3. Разделение данных на обучающую и тестовую выборки.
4. Создание и обучение модели логистической регрессии.
5. Оценка модели.

Образец кода в example.csv

Данные состояли из 768 записей без пропущенных значений. После разделения на обучающую (537 записей) и тестовую выборки (231 запись), модель показала точность 74% на тестовой выборке. Матрица ошибок показала хорошие результаты с 120 True Negative и 50 True Positive, хотя ошибки (31 False Positive и 30 False Negative) также присутствуют. ROC-кривая с AUC = 0.80 указывает на хорошую способность модели различать два класса. Таким образом, программа успешно выполнила задачи по обработке данных, построению модели и оценке её качества.

Ответить на вопросы:

1. Что такое логистическая регрессия и в каких случаях она применяется?

Логистическая регрессия — это метод статистического анализа, используемый для решения задач классификации. Она применяется, когда зависимая переменная является бинарной (например, «да» или «нет», «1» или «0»). Модель предсказывает вероятность принадлежности объекта к одному из двух классов.

1. Какова основная цель логистической регрессии?

Основная цель логистической регрессии — предсказать вероятность, с которой объект принадлежит к определённому классу, и разделить данные на два класса на основе этой вероятности.

1. В чем разница между логистической регрессией и линейной регрессией?

Линейная регрессия используется для прогнозирования непрерывных значений, таких как цена или рост.

Логистическая регрессия используется для задач классификации, где результат представляет собой вероятность принадлежности к определённому классу. Она использует логистическую функцию (сигмоиду), чтобы преобразовать линейную комбинацию переменных в значения от 0 до 1.

1. Какие предпосылки необходимо проверить перед применением логистической регрессии?

Линейная связь между независимыми переменными и логарифмом шансов (log-odds).

Отсутствие сильной мультиколлинеарности между независимыми переменными.

Независимость наблюдений.

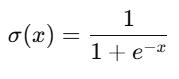
Наличие достаточного количества данных для каждого класса.

1. Как интерпретировать коэффициенты логистической регрессии?

Коэффициенты логистической регрессии показывают изменение логарифма шансов (log-odds) при изменении соответствующей независимой переменной на одну единицу. Экспонента коэффициента (eβ) интерпретируется как отношение шансов (odds ratio).

1. Что такое функция логистической (сигмоидной) функции и как она используется в логистической регрессии?

Логистическая функция (сигмоида) — это математическая функция, преобразующая любое значение в диапазон от 0 до 1. Формула:



В логистической регрессии она используется для преобразования линейной комбинации входных переменных в вероятность принадлежности к классу.

1. Каковы основные метрики для оценки качества модели логистической регрессии?

* Точность (Accuracy): Доля правильно классифицированных наблюдений.
* Матрица ошибок (Confusion Matrix): Показывает количество TP (True Positive), TN (True Negative), FP (False Positive) и FN (False Negative).
* Precision (Точность): Доля истинных положительных результатов среди всех предсказанных положительных.
* Recall (Полнота): Доля истинных положительных результатов среди всех реальных положительных.
* F1-score: Гармоническое среднее Precision и Recall.
* ROC-кривая и AUC (Площадь под ROC-кривой): Оценивают способность модели различать классы.