Министерство науки и высшего образования РФ

ФГБОУ ВО «Удмуртский государственный университет»

Институт математики, информационных технологий и физики

Кафедра теоретических основ информатики

Отчет по дисциплине «Машинное обучение»

Лабораторная работа №2 “Задача бинарной классификации c использованием классификатора KNN”

Выполнил:

студент группы ОМ-02.04.02.04-11

Данилов Владимир Александрович

Принял:

К.ф-м.н.

Федоров Дмитрий Леонидович

Ижевск, 2023 г.

**Содержание**

[Постановка задачи 3](#__RefHeading___1)

[Краткая теория 3](#__RefHeading___2)

[Результаты тестирования модели 6](#__RefHeading___3)

[Анализ результатов и выводы 7](#__RefHeading___4)

# Постановка задачи

Требуется написать программу, решающую задачу бинарной классификации. Для этого требуется выполнить следующее:

1. Загрузить набор данных по анализу и прогнозированию инфаркту как датафрейм библиотеки pandas. Источник данных: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>.
2. Выполнить построение модели логистической регрессии, используя библиотеку sklearn.
3. Провести перекрестную проверку полученной модели.
4. Построить ROC-кривую для получения классификатора

# Краткая теория

Классификатор k ближайших соседей (k-nearest neighbors classifier, KNN) это один из простейших алгоритмов машинного обучения, который используется для классификации объектов.

Описание работы алгоритма KNN:

1. Имеется набор данных, состоящий из объектов с уже известными метками классов.

2. Для классификации нового неизвестного объекта алгоритм находит k ближайших к нему объектов из обучающей выборки.

3. Определяется класс нового объекта по принципу "голосования" (большинство). Например, если среди k ближайших соседей большинство объектов относятся к классу A, то новый объект также относится к классу A.

4. Алгоритм возвращает предсказанный класс нового объекта.

* Преимущества KNN:
* Простота реализации и понимания;
* Не требует предварительного обучения модели;
* Может работать с различными типами данных.

Недостатки KNN:

* Значительное вычислительное время при больших наборах данных;
* Чувствительность к выбросам и несбалансированным данным;
* Чувствительность к выбору гиперпараметра k, который определяет количество ближайших соседей.

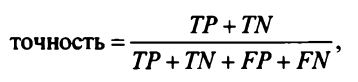
**Перекрестная проверка модели** – это метод оценки производительности модели машинного обучения, который помогает оценить ее предсказательную способность на независимых данных.

В перекрестной проверке модель разделяется на несколько групп или фолдов. Затем модель обучается на некоторых фолдах и проверяется на оставшихся фолдах. Процесс повторяется несколько раз, пока каждый фолд не будет использован как проверочный набор данных. В конце полученные показатели производительности модели усредняются.

Перекрестная проверка позволяет уменьшить вероятность переобучения модели и несмещенно оценить ее производительность на реальных данных. Она также помогает выбрать подходящие гиперпараметры модели, такие как скорость обучения или глубина дерева в деревьях принятия решений.

Наиболее распространенным методом перекрестной проверки является k-fold кросс-валидация, где модель делится на k групп, из которых одна группа используется для проверки, а остальные - для обучения. Этот процесс повторяется k раз, чтобы каждая группа была использована в качестве проверочного набора данных.

*Точность* является общепринятым метрическим показателем результативности.



где:

*ТР* – количество истинноположительных исходов; наблюдения, которые являются частью положительного класса (имеет заболевание, приобретен товар

и т. д.) и которые мы предсказали правильно;

*TN* – количество истинноотрицательных исходов; наблюдения, которые являются частью отрицательного класса (не имеет этого заболевания, не преобретен

товар и т. д.) и которые мы предсказали правильно;

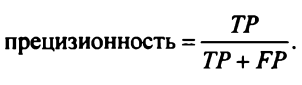
*FP* – количество ложноположительных исходов, также называется ошибкой

1-го рода; наблюдения, предсказанные как часть положительного класса, которые на самом деле являются частью отрицательного класса;

*FN* – количество ложноотрицательных исходов, также называется ошибкой

2-го рода; наблюдения, предсказанные как часть отрицательного класса, которые на самом деле являются частью положительного класса.

*Прецизионность* – это доля каждого наблюдения, предсказанного положительно, которое на самом деле положительно. Эту меру можно представить, как измерительный шум в наших предсказаниях, т. е. когда мы предсказываем что-то положительное, насколько вероятно, что мы будем правы. Модели с высокой прецизионностью пессимистичны, т. е. они предсказывают наблюдение как принадлежащее положительному классу, только когда они очень уверены в этом. Формально прецизионность рассчитывается по



*Полнота* – это доля каждого положительного наблюдения, которое по-настоящему положительно. Полнота измеряет способность модели идентифицировать наблюдение положительного класса. Модели с высокой полнотой оптимистичны, т. е. они имеют низкую планку для предсказания, что наблюдение находится в положительном

# Результаты тестирования модели

Натренировать модель:

[[-0.05395622 -0.76912513 0.84304206 -0.3189122 -0.21817738 0.01343313

0.23229389 0.50455195 -0.45035284 -0.59894956 0.3440393 -0.75250858

-0.53093014]]

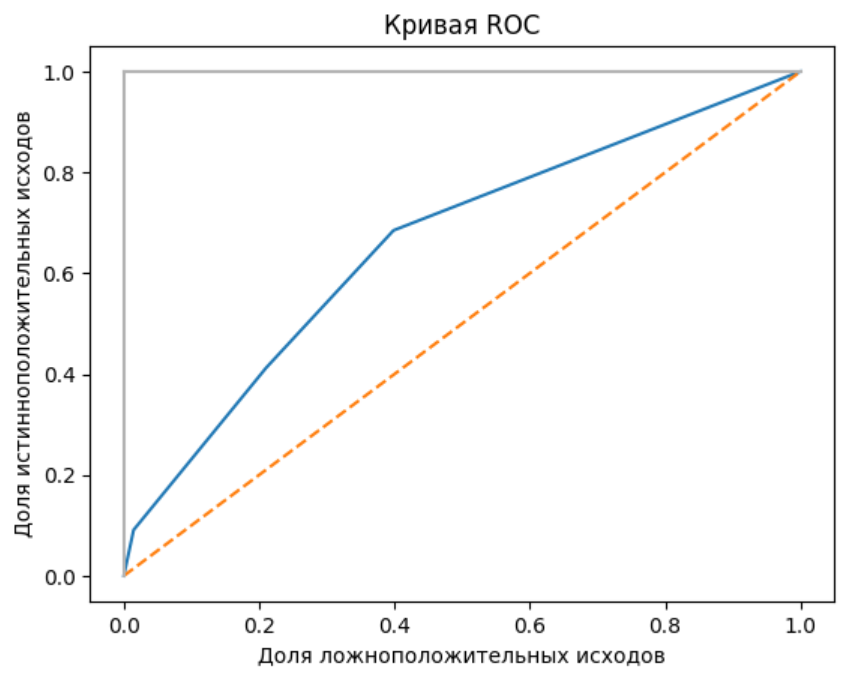
Перекрестная проверка: 0.8418279569892473

Точность: [0.62295082 0.62295082 0.67213115 0.7 0.61666667]

Полнота: [0.72727273 0.72727273 0.72727273 0.78787879 0.66666667]

Прецизионность: [0.63157895 0.63157895 0.68571429 0.7027027 0.64705882]

AUC: 0.6598814229249013



# Анализ результатов и выводы

Значение по всем показателям являются средними. Это означат, что построенная модель не может точно описывать задачу бинарной классификации для предоставленного набора данных, например, возьмем показатель AUC = 0.6598814229249013, данное значение является средним и означает, что модель обладает не самой худшей прогностической силой, но и с другой стороны означает, что модель не может предсказать стопроцентный прогноз результата.

Сравнивая значения, полученные в первой лабораторной работе, где показатель AUC = 0.7523056653491436, можно сделать утверждение, что первая модель, использующая логистическую регрессию, дает результат лучше, чем модель с классификатором k ближайших соседей.