Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

*Факультет программной инженерии и компьютерной техники*

Системы искусственного интеллекта

Лабораторная работа №4

Метод k-ближайших соседей

Группа: P3324

Выполнил: Маликов Глеб Игоревич

Преподаватель:

Королёва Юлия Александровна

Санкт-Петербург

2024г.

**Оглавление**

[Введение 3](#_Toc178333027)

[Описание метода 4](#_Toc178333028)

[Реализация метода 5](#_Toc178333029)

[Результаты выполнения 6](#_Toc178333030)

[Обработка значений 6](#_Toc178333031)

[Визуализация статистики по датасету 6](#_Toc178333032)

[Корреляционная матрица 9](#_Toc178333033)

[3D-визуализация признаков 10](#_Toc178333034)

[Модель 1: Признаки случайно отбираются 10](#_Toc178333035)

[Модель 2: Фиксированный набор признаков 10](#_Toc178333036)

[Примеры использования метода 12](#_Toc178333037)

# Введение

В данной лабораторной работе рассматривается задача классификации пациентов на основе признаков, связанных с диабетом. Цель работы заключается в применении метода k-ближайших соседей (k-NN) для построения моделей, способных предсказывать вероятность наличия диабета у пациента. Для достижения этой цели выполнены следующие задачи:

1. Проведена предварительная обработка данных, включая обработку отсутствующих и нулевых значений, кодирование категориальных признаков и масштабирование данных.
2. Проведена визуализация статистики по датасету, включая построение гистограмм, корреляционной матрицы и 3D-визуализацию признаков.
3. Реализован алгоритм метода k-ближайших соседей без использования сторонних библиотек, кроме NumPy и Pandas.
4. Построены две модели k-NN с различными наборами признаков: случайно отобранными и фиксированным набором признаков.
5. Проведена оценка моделей на тестовом наборе данных при различных значениях параметра k и построены матрицы ошибок.

# Описание метода

Метод k-ближайших соседей (KNN) — это простой, но мощный алгоритм для решения задач классификации и регрессии, основанный на предположении, что объекты, которые находятся близко друг к другу в пространстве признаков, имеют схожие свойства.

Для того чтобы классифицировать или предсказать значение нового объекта, алгоритм находит k объектов (соседей) в тренировочной выборке, которые находятся к нему ближе всего, и на основании их меток классов (или значений) делает вывод.

Близость между объектами обычно измеряется с использованием различных метрик расстояния, таких как евклидово расстояние, манхэттенское расстояние или косинусное расстояние.

Ключевой параметр в KNN — это количество ближайших соседей k, которое определяется заранее. Если выбрать слишком маленькое k, алгоритм может быть чувствителен к шуму и отдельным выбросам, так как решение будет основываться на слишком малом количестве объектов. Слишком большое k, наоборот, может привести к "размыванию" локальной информации и учитывать слишком много далеко расположенных объектов.

# Реализация метода

В реализации метода k-NN использовались библиотеки NumPy и Pandas для обработки данных. Основные шаги реализации включают:

* Вычисление евклидового расстояния между точкой тестирования и всеми точками обучающей выборки.
* Определение k ближайших соседей на основе минимальных расстояний.
* Определение класса среди классов ближайших соседей.
* Построение матрицы ошибок для оценки точности модели при различных значениях k.

Код реализации метода представлен ниже:

def euclidean\_distance(point1, point2):  
 return np.sqrt(np.sum((point1 - point2) \*\* 2))  
def k\_nearest\_neighbors(train\_data, train\_labels, test\_point, k=3):  
 train\_data = np.array(train\_data)  
 train\_labels = np.array(train\_labels)  
   
 distances = []  
 for i in range(len(train\_data)):  
 distance = euclidean\_distance(train\_data[i], test\_point)  
 distances.append((distance, train\_labels[i]))  
   
 nearest\_indices = np.argpartition([distance for distance, \_ in distances], k)[:k]  
 nearest\_labels = [int(train\_labels[i]) for i in nearest\_indices]  
   
 most\_common = np.bincount(nearest\_labels).argmax()  
 return most\_common

# Результаты выполнения

## Обработка значений

Были удалены строки с нулевыми значениями для столбцов Glucose, BloodPressure, SkinThickness и BMI, так как пациенты не могут иметь такие нулевые значения.

Zero values per column:

Glucose 5

BloodPressure 35

SkinThickness 227

BMI 11

dtype: int64

Deleted rows: 236

Также была проведена нормализация значений по формуле:

## Визуализация статистики по датасету

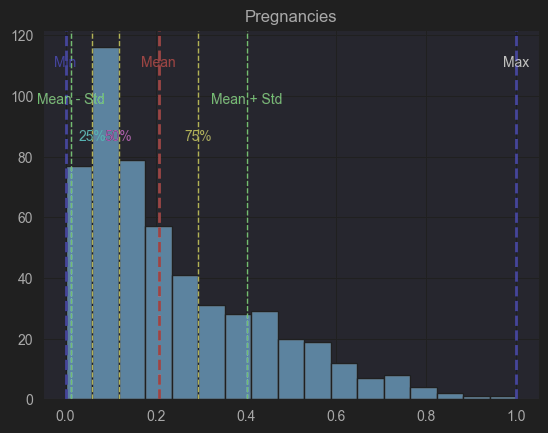


График 1 - Статистика "Pregnancies"

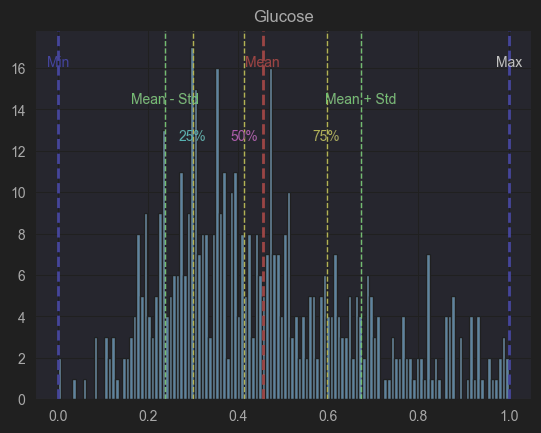


График 2 - Статистика "Glucose"

A screen shot of a graph

Description automatically generated

График 3 - Статистика "BloodPressure"

A graph of a graph

Description automatically generated with medium confidence

График 4 - Статистика "SkinThickness"

A screenshot of a graph

Description automatically generated

График 5 - Статистика "Insulin"

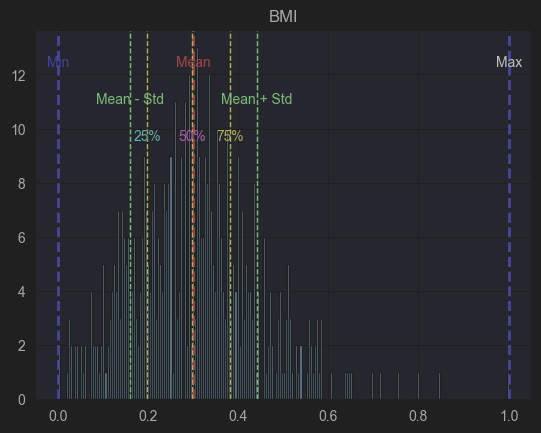


График 6 - Статистика "BMI"

A graph with numbers and text

Description automatically generated with medium confidence

График 7 - Статистика "Pedigree"

A graph of a number of people

Description automatically generated with medium confidence

График 8 - Статистика "Age"

## Корреляционная матрица

Для выбора подходящих значений для модели с фиксированными признаками была построена корреляционная матрица.

A screenshot of a computer screen

Description automatically generated

График 9 - Корреляционная матрица

Таким образом можно обнаружить что признаки с наибольшей положительной корреляцией с признаком “Outcome” являются признаки “Glucose”, “Age” и “BMI”

## 3D-визуализация признаков

A screen shot of a graph

Description automatically generated

График 10 - 3D Визуализация признаков "Glucose", "BMI" и "Age"

## Модель 1: Признаки случайно отбираются

В этой модели случайно были выбраны признаки “BloodPressure”, “SkinThickness” и “Pedigree”. Полученные матрицы ошибок и их точность показаны ниже:

Confusion matrix for k=3:

[[50 23]

[17 16]]

0.6226415094339622

Confusion matrix for k=5:

[[58 15]

[16 17]]

0.7075471698113207

Confusion matrix for k=10:

[[63 10]

[23 10]]

0.6886792452830188

## Модель 2: Фиксированный набор признаков

В этой модели используются признаки, показавшие наибольшую положительную корреляцию с “Outcome”, то есть признаки “Glucose”, “Age” и “BMI”.

Confusion matrix for k=3:

[[58 15]

[12 21]]

0.7452830188679245

Confusion matrix for k=5:

[[60 13]

[12 21]]

0.7641509433962265

Confusion matrix for k=10:

[[66 7]

[14 19]]

0.8018867924528302

Из представленных результатов видно, что фиксированный набор признаков (Glucose, BMI, Age) демонстрирует более высокую точность классификации по сравнению с случайно отобранными признаками. Наилучшую точность достигает модель с k=10, что составляет 80.19%.

# Примеры использования метода

Метод k-ближайших соседей широко применяется в различных областях благодаря своей простоте и эффективности. Примеры ситуаций, где метод k-NN может быть полезен:

1. Медицинская диагностика: как показано в данной лабораторной работе, метод k-NN может использоваться для классификации пациентов на основе медицинских показателей, помогая врачам в диагностике заболеваний, таких как диабет.
2. Рекомендательные системы: В системах рекомендаций, например, для онлайн-магазинов или стриминговых сервисов, метод k-NN помогает предлагать пользователям товары или контент на основе предпочтений схожих пользователей.

Причины выбора метода k-NN:

* Простота реализации: Метод легко реализовать и интерпретировать, что делает его подходящим для начальных этапов анализа данных.
* Не требуется обучение модели: k-NN является алгоритмом, не требующим предварительного обучения, что упрощает его применение на новых данных.
* Эффективен при небольших объемах данных: для задач с относительно небольшими наборами данных метод демонстрирует хорошую производительность.

Однако следует учитывать, что метод k-NN может быть менее эффективен при работе с большими и высоко размерными наборами данных из-за высокой вычислительной сложности и чувствительности к выбору признаков.