Определение вида активности человека

Хачатурова Ирина, Леванова Марина, 331 гр

Данные

Эксперименты проводились с группой из 30 добровольцев в возрасте от 19 до 48 лет. Каждый человек выполнял шесть действий (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) с прикреплённым к поясу смартфоном (Samsung Galaxy S II). С помощью встроенного в смартфон акселерометра и гироскопа мы фиксировали трёхмерное линейное ускорение и трёхмерную угловую скорость с постоянной частотой 50 Гц. Эксперименты были записаны на видео для ручной маркировки данных. Полученный набор данных был случайным образом разделён на две части: 70% добровольцев были отобраны для создания обучающих данных, а 30% — для создания тестовых данных.

В датасете каждый признак - это какая-то операция над вектором из 128 чисел, которые соответствуют замерам определенного человека, занятого какой-то активностью на протяжении 2.56 секунд.

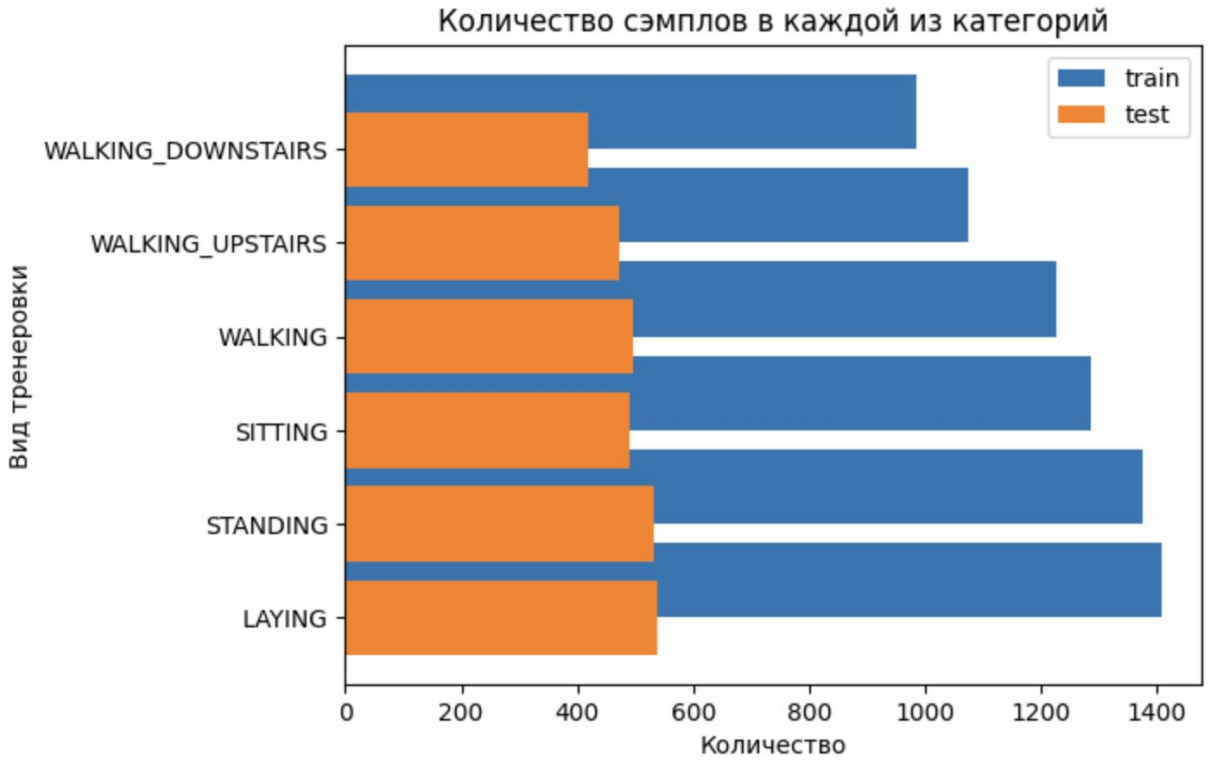
Обзор данных

В датасете 563 признака, включая целевую переменную. Мы выбрали для нашей задачи следующие:

- 'tBodyAccMag-mean()', # усредненная величина ускорения тела
- 'tBodyGyroJerk-mad()-X', # медианная величина рывка тела по оси X
- 'tGravityAcc-min()-X', # минимум гравитационной составляющей ускорения по оси X
- 'tBodyAcc-max()-X', # максимальная величина ускорения тела по оси X
- 'fBodyAcc-bandsEnergy()-1,8.2', # энергия ускорения тела в интервале частоты
- 'angle(X,gravityMean)', # угол между осью X и усредненной гравитационной составляющей ускорения
- 'angle(Y,gravityMean)', # угол между осью Y и усредненной гравитационной составляющей ускорения
- 'angle(Z,gravityMean)', # угол между осью Z и усредненной гравитационной составляющей ускорения,
- 'fBodyAcc-skewness()-X', # асимметричность частоты ускорения тела по оси X
- 'subject', # номер испытуемого
- 'Activity', # название вида деятельности (целевая переменная)

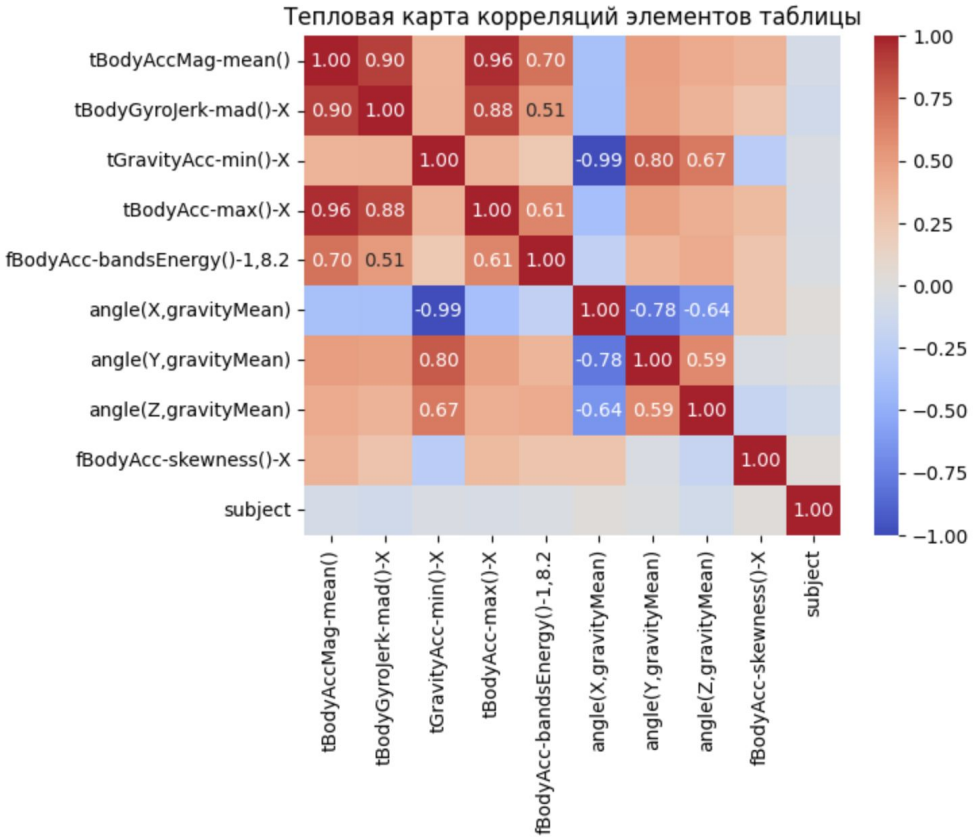
Предварительный анализ данных

Визуализация количество сэмплов в каждой из категорий отдельно для train и test сетов. Порядок категорий слева и справа одинаков.



Предварительный анализ данных

Визуализация тепловой карты корреляций



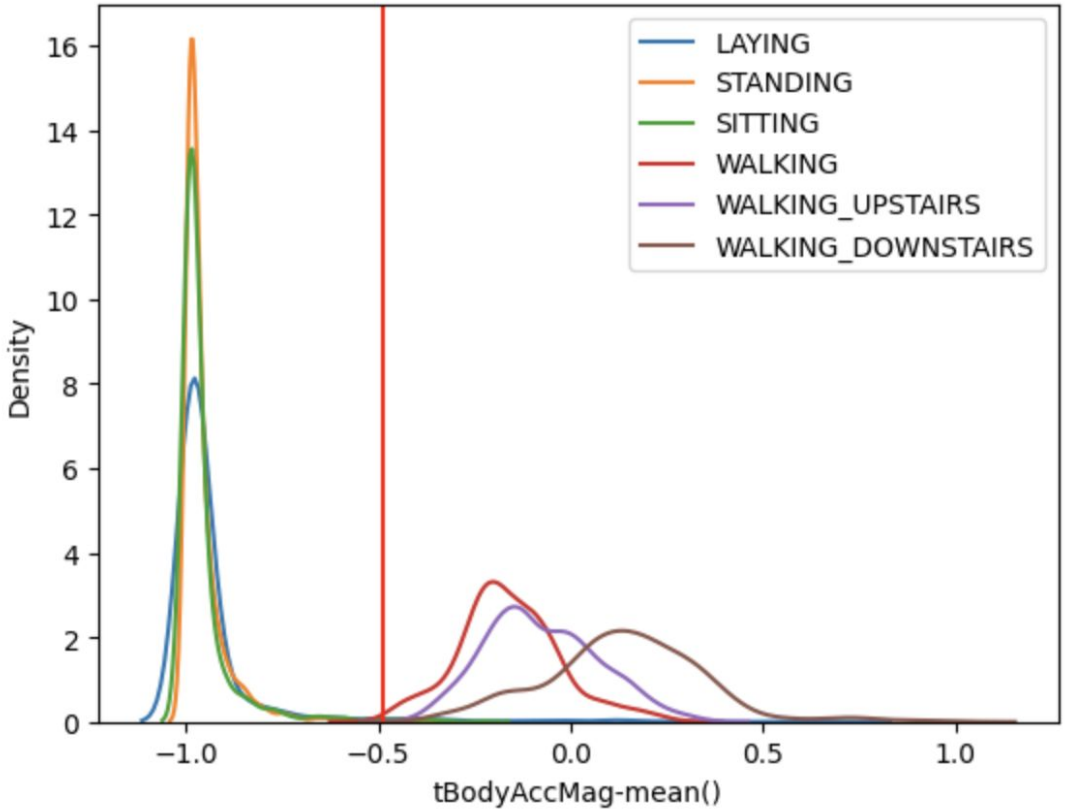
Предварительный анализ данных

Визуализация плотностей распределения `tBodyAccMag-mean()` для каждой из активностей.

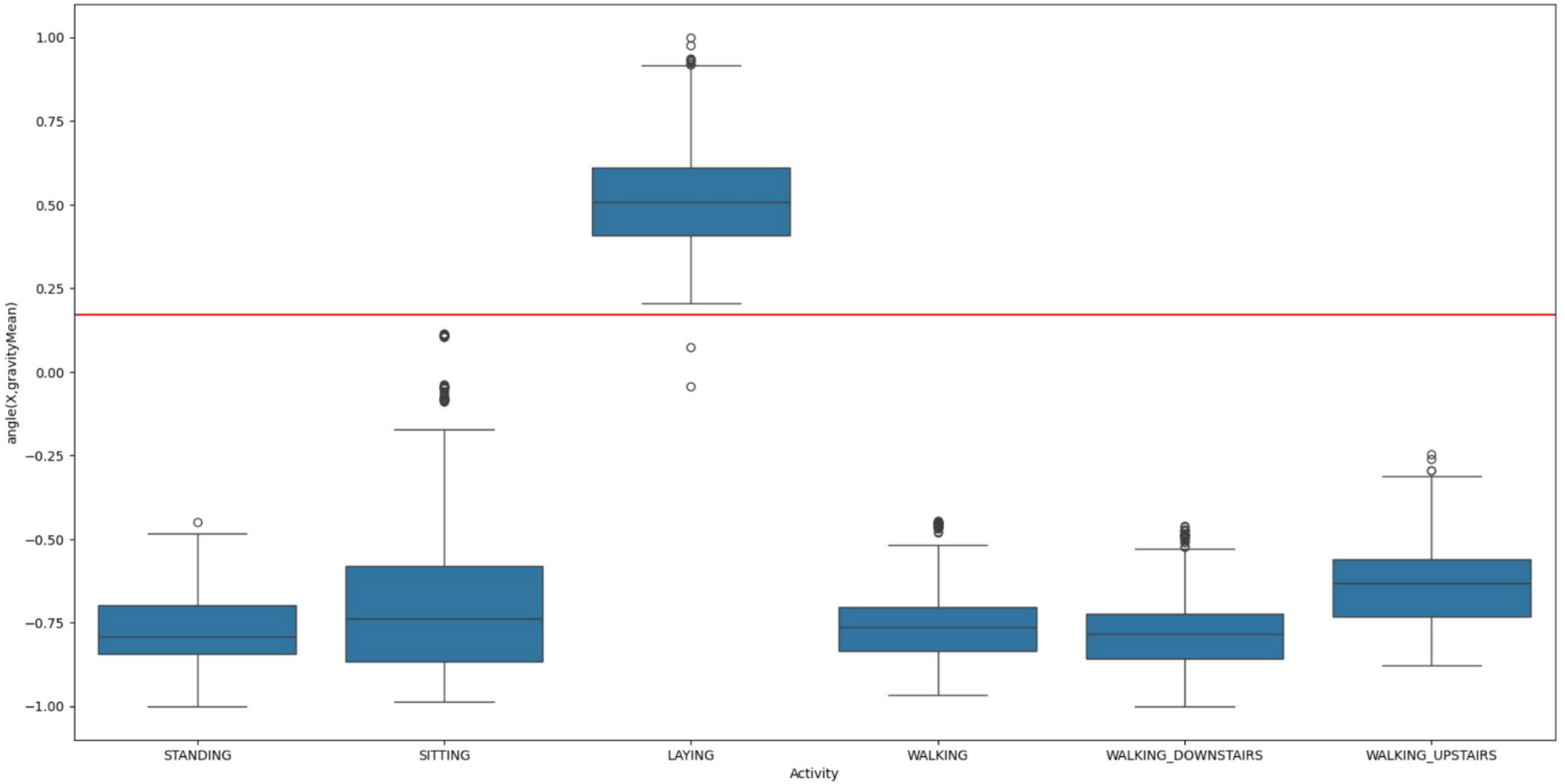
Можно поделить все активности на две группы - проведем вертикальную линию, которая разделяет эти группы.

Первая группа - статические активности, вторая - динамическиие.

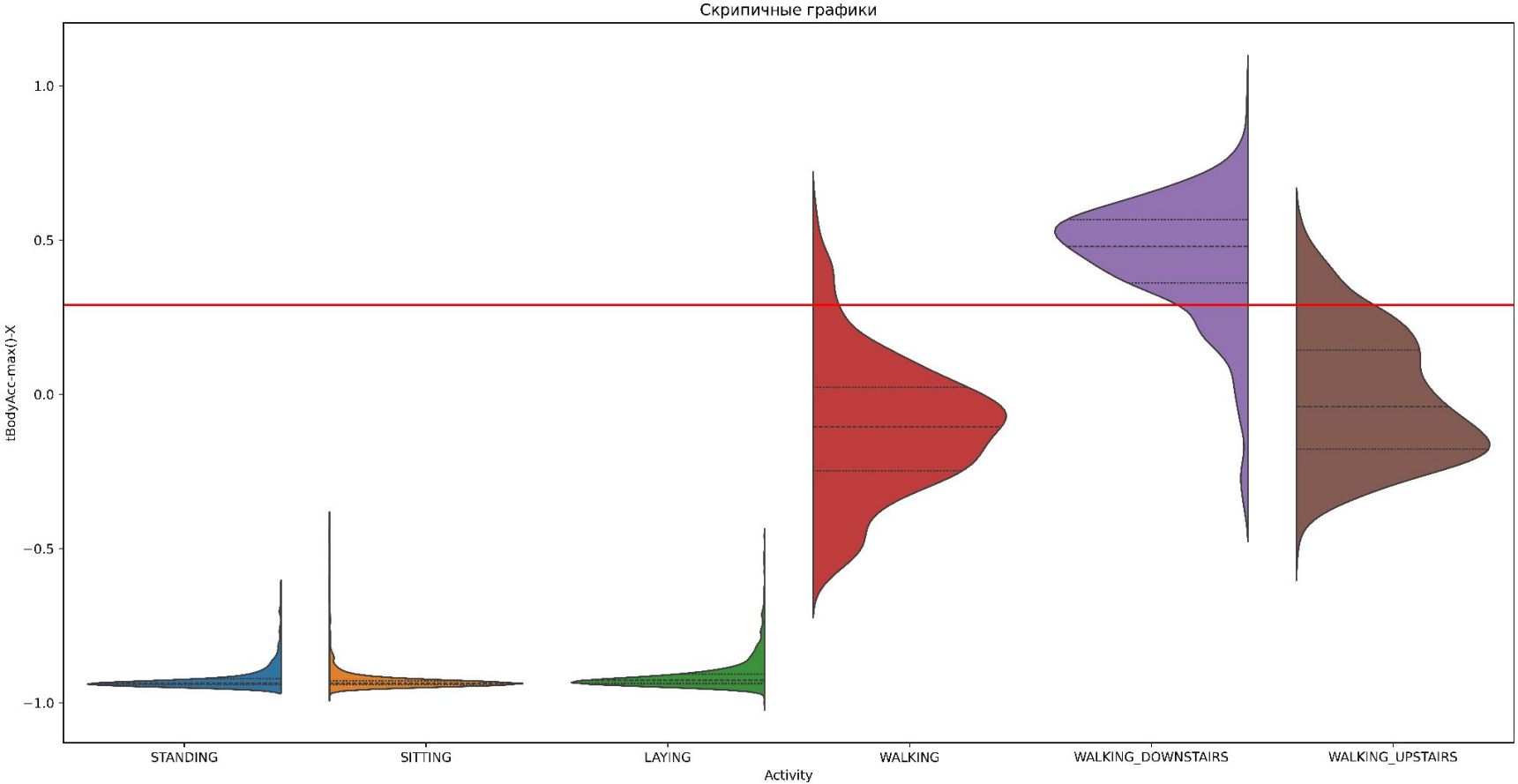
*tBodyAccMag-mean() - усредненная величина ускорения тела



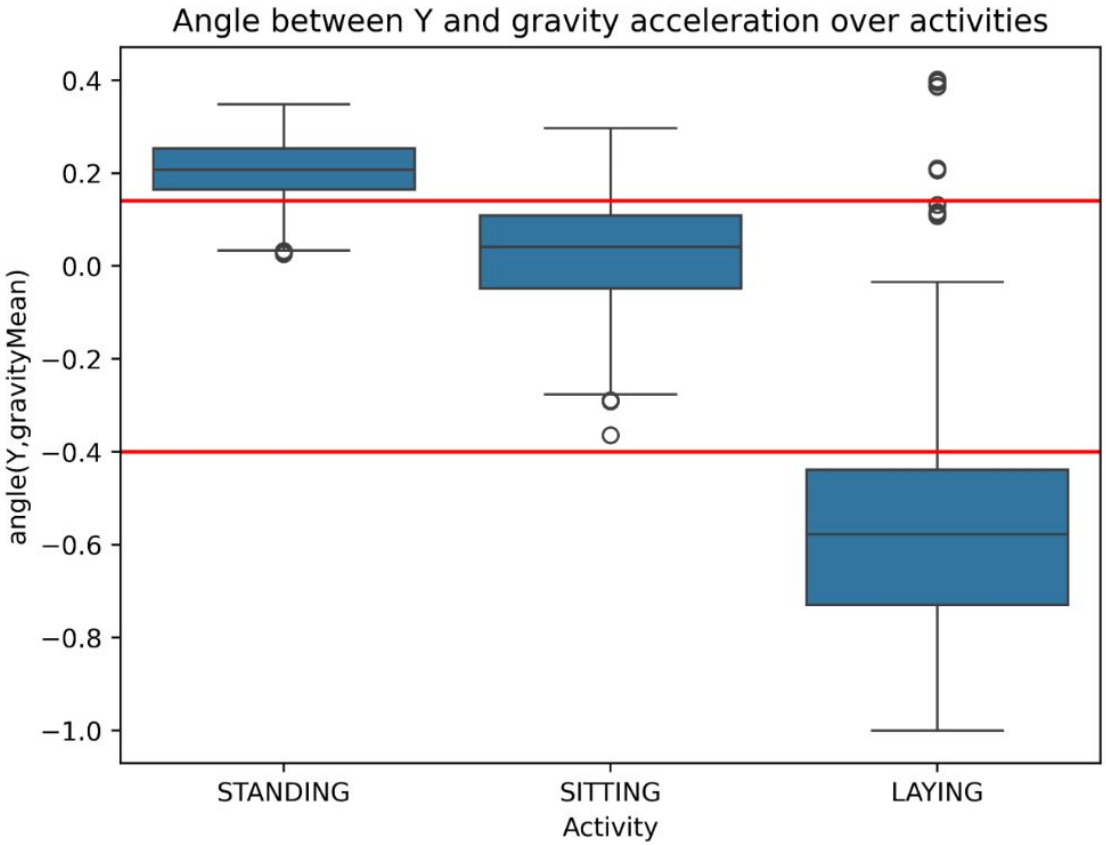
Предварительный анализ данных



Предварительный анализ данных



Предварительный анализ данных



Метод knn

Формальное определение:

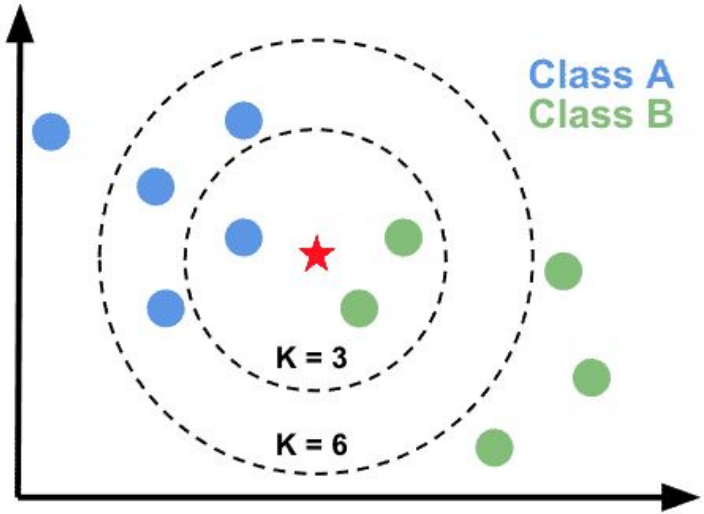
$$\hat{y} = \text{argmax}_k \sum_{x_i \in N_k(x)} I(y_i = k)$$

Параметры:

- Расстояние евклидово

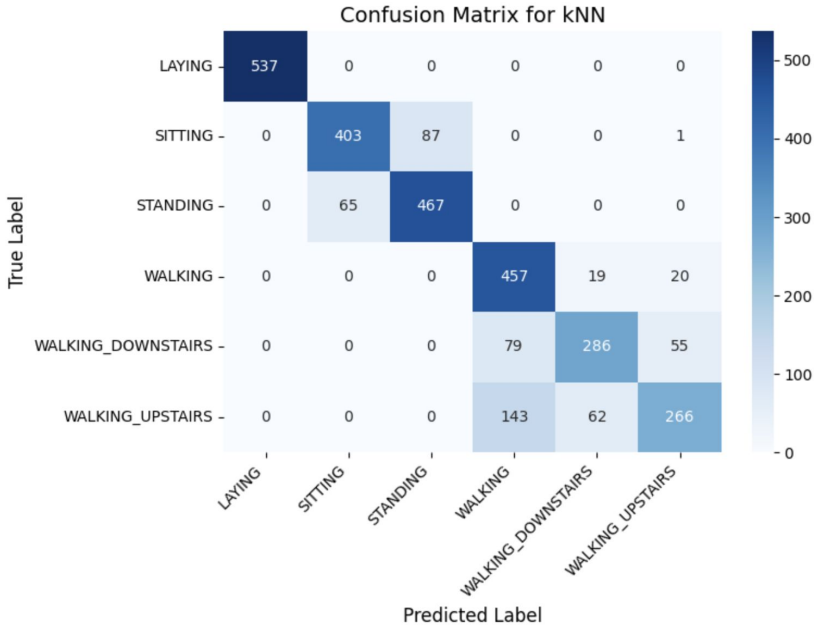
$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

- k = 18 (подобрано через GridSearch)

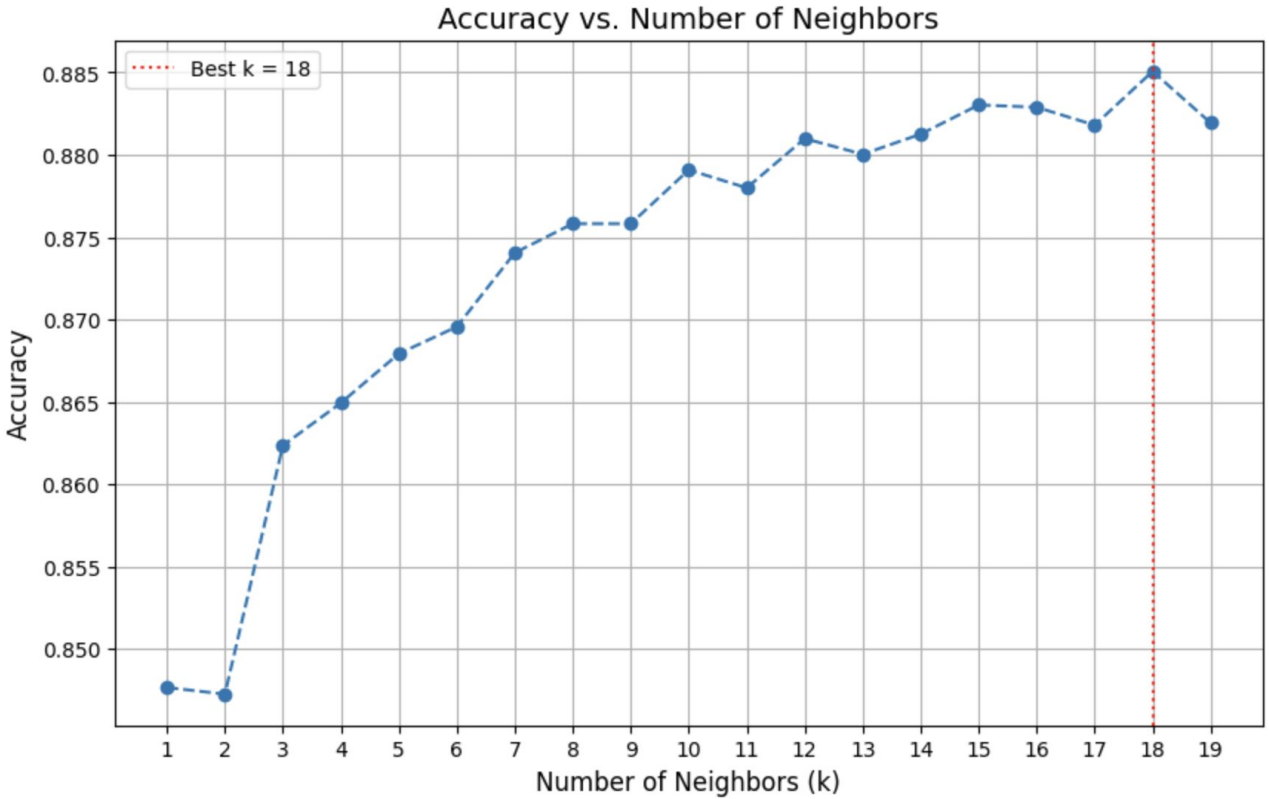


Метод knn. Результаты

```
Оптимальное количество соседей: 18

Оценка качества kNN:
Accuracy: 0.8910756701730573
Precision (macro): 0.9019756091487411
Recall (macro): 0.8848797786869794
F1-score (macro): 0.8870668795031129

Classification Report:
                    precision    recall  f1-score   support

            LAYING       1.00      0.95      0.97       537
           SITTING       0.92      0.80      0.85       491
          STANDING       0.82      0.95      0.88       532
           WALKING       0.81      0.99      0.89       496
WALKING_DOWNSTAIRS       0.98      0.71      0.83       420
  WALKING_UPSTAIRS       0.89      0.90      0.89       471

          accuracy                           0.89      2947
         macro avg       0.90      0.88      0.89      2947
      weighted avg       0.90      0.89      0.89      2947
```



Метод knn. Результаты



Метод случайного леса

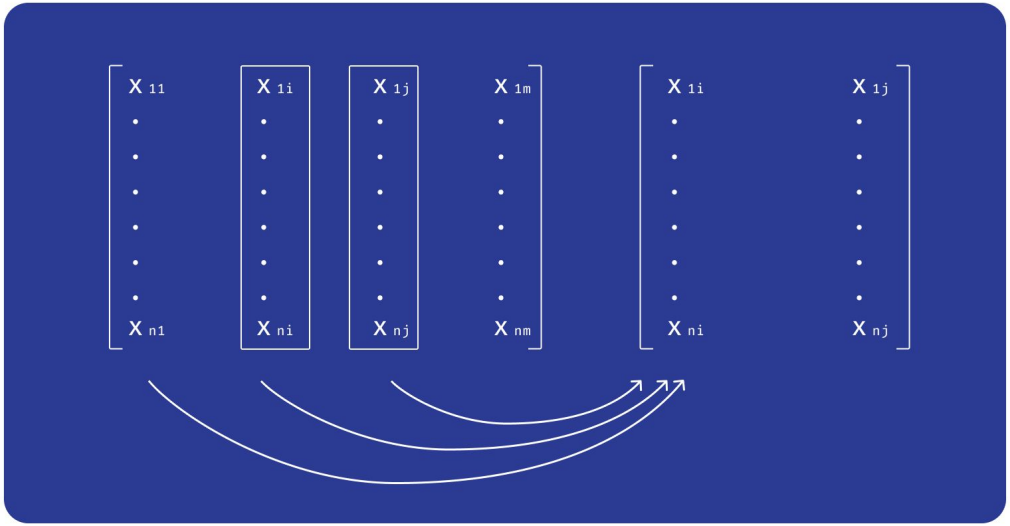
Случайный лес генерирует множество различных независимых друг от друга деревьев слегка разными способами (берёт разные подвыборки, разные признаки), а на основании их ответов формирует итоговое решение. Алгоритм случайного леса усредняет ответы всех деревьев (в задаче регрессии) или выбирает голосованием (в классификации) тот ответ, который большинство деревьев в лесу считает правильным.

Деревья обучаются параллельно — независимо друг от друга. Обучение каждого дерева не зависит от результатов других.

Метод случайного леса сочетает в себе два подхода: метод случайных подпространств и бэггинг.

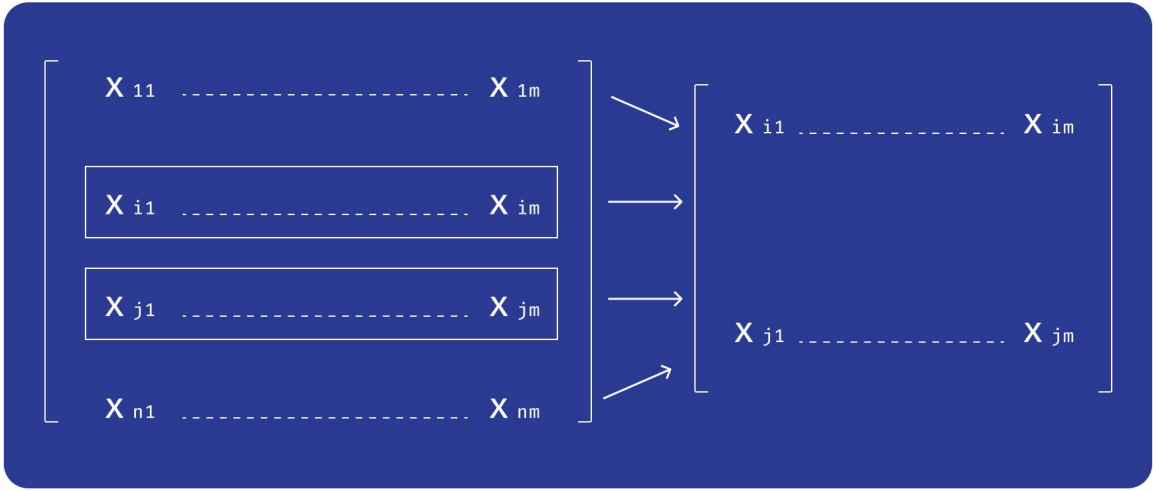
Метод случайных подпространств

Один из способов сделать базовые модели (деревья в нашем случае) как можно более разными - обучать их на случайно выбранных подвыборках признаков объектов.

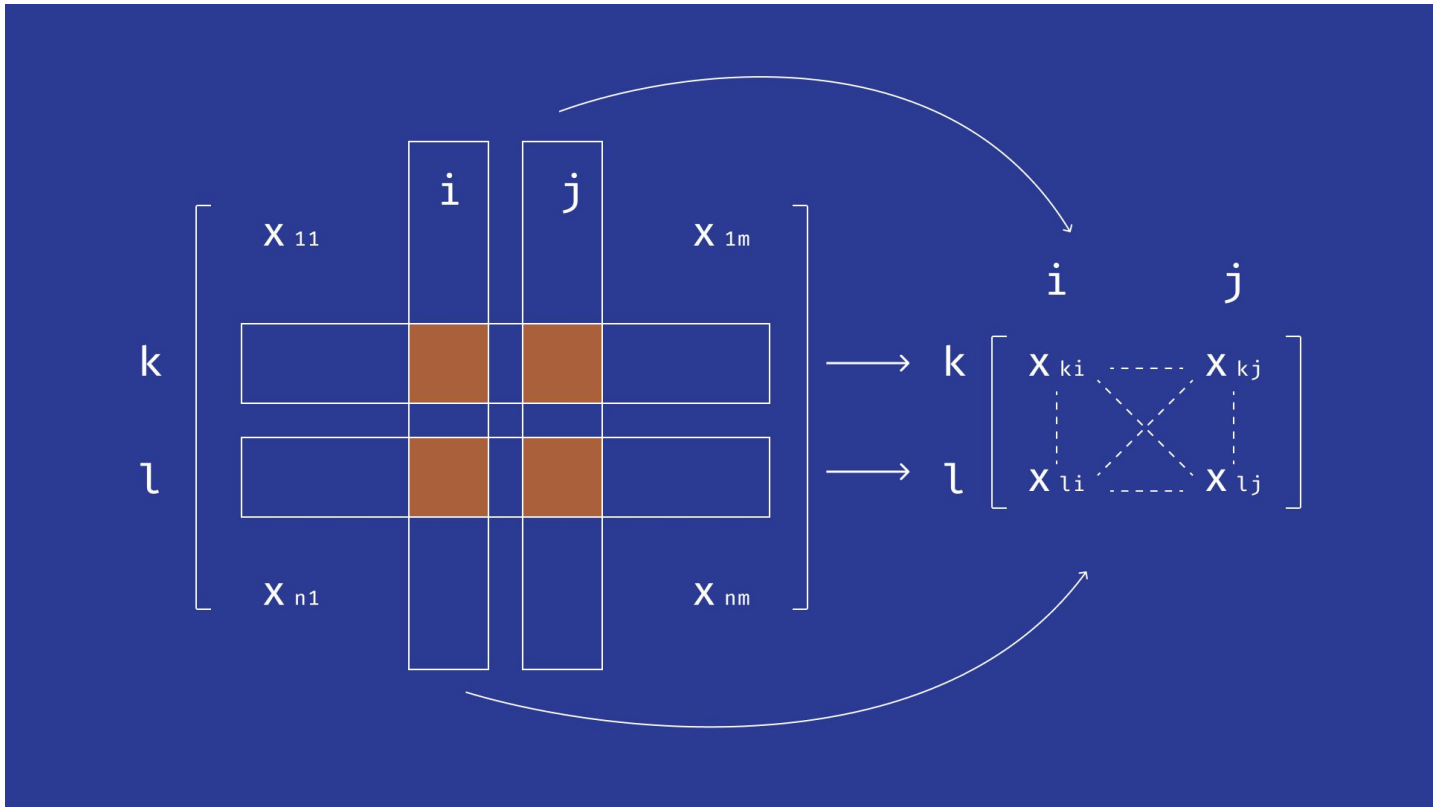


Бэггинг

Бэггинг предполагает выбор случайных подвыборок самих объектов. Случайная подвыборка формируется с повторениями (то есть если мы выбрали некоторый объект из выборки, мы можем выбрать его еще раз. При таком подходе в подвыборку в среднем попадает около 63% объектов исходной выборки).

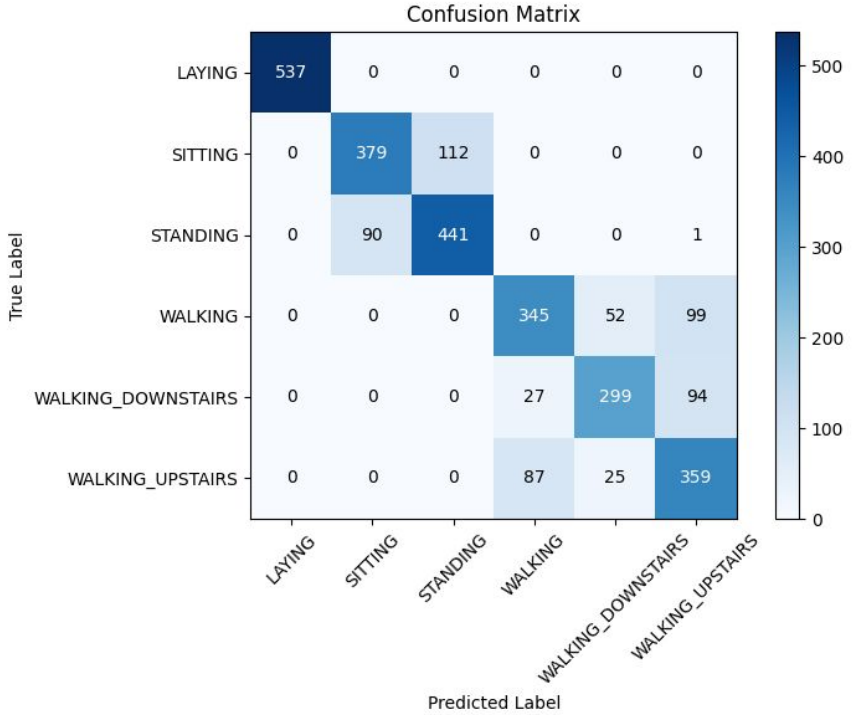


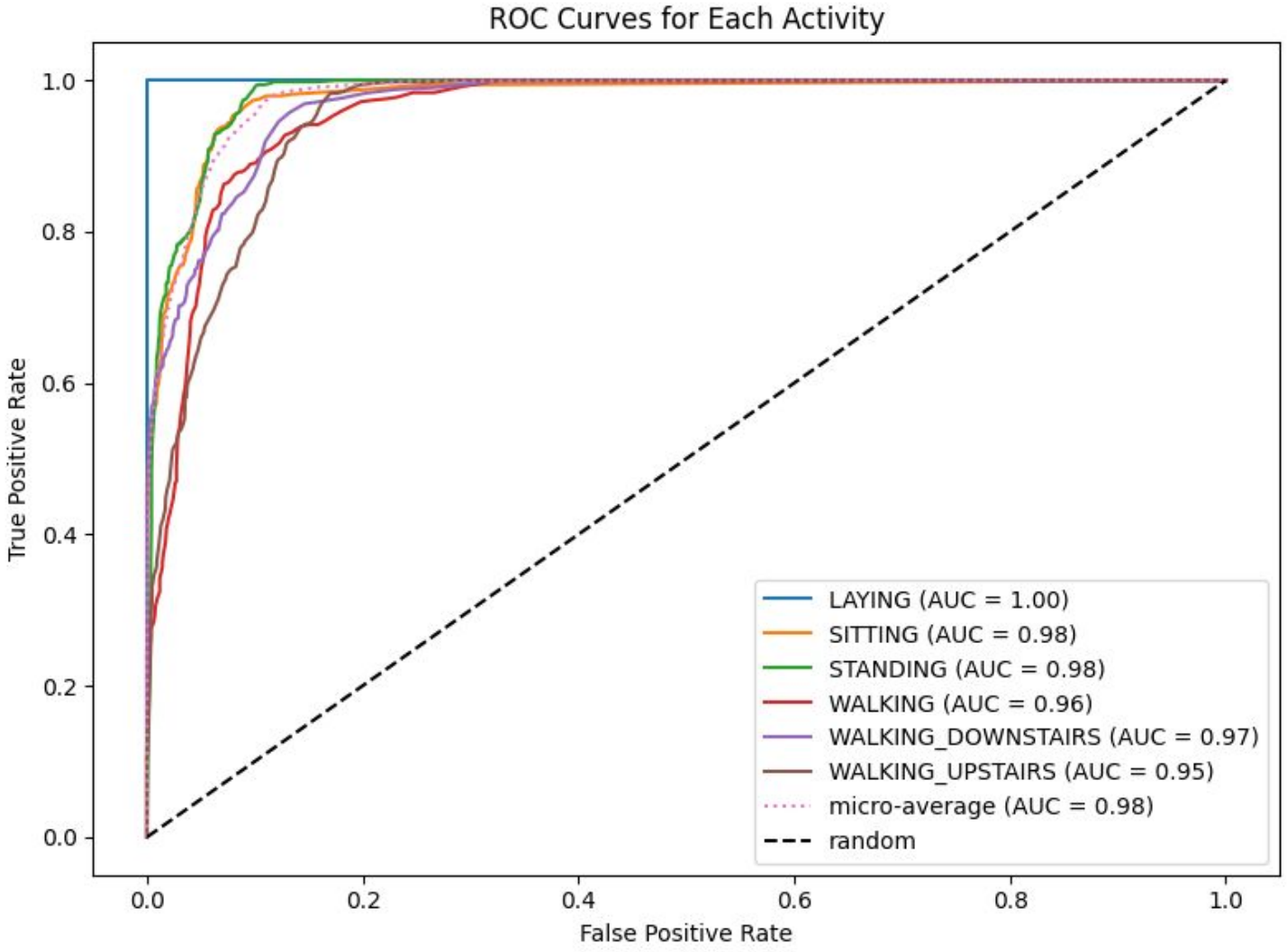
Случайный лес



Метод случайного леса. Результаты

	precision	recall	f1-score	support
LAYING	1.00	1.00	1.00	537
SITTING	0.81	0.77	0.79	491
STANDING	0.80	0.83	0.81	532
WALKING	0.75	0.70	0.72	496
WALKING_DOWNSTAIRS	0.80	0.71	0.75	420
WALKING_UPSTAIRS	0.65	0.76	0.70	471
accuracy			0.80	2947
macro avg	0.80	0.80	0.80	2947
weighted avg	0.80	0.80	0.80	2947




ROC Curves for Each Activity
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LAYING (AUC = 1.00)
SITTING (AUC = 0.98)
STANDING (AUC = 0.98)
WALKING (AUC = 0.96)
WALKING_DOWNSTAIRS (AUC = 0.97)
WALKING_UPSTAIRS (AUC = 0.95)
micro-average (AUC = 0.98)
random