

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И  
РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра компьютерных систем в управлении и проектировании (КСУП)

# РАЗРАБОТКА ПРОГРАММЫ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДИАГНОСТИКИ БОЛЕЗНИ ПАРКИНСОНА НА НЕСБАЛАНСИРОВАННОМ НАБОРЕ ДАННЫХ HANDPD

Выполнил: студент 4 курса, группы 589-3  
Пахомов Максим Владимирович

Руководитель работы: доцент кафедры КСУП, к.т.н.  
Бардамова Марина Борисовна

Томск 2023

## Цель и задачи

Цель: разработать программу для выявления болезни Паркинсона (БП) по оцифрованным изображениям рукописных рисунков и выполнить проверку её эффективности на наборе данных HandPD.

Задачи:

- изучение предметной области и исследование набора данных HandPD;
- разработка программы для выявления БП по оцифрованным изображениям;
- апробация программы на наборе данных HandPD.

## Актуальность

БП — это дегенеративное заболевание нервной системы, которое приводит к нарушению координации движений, снижению когнитивных функций и другим серьезным проблемам. Согласно данным Всемирной организации здравоохранения, более 10 миллионов человек страдают от болезни Паркинсона в мире.

Существующие в настоящий момент методы диагностики БП неэффективны на ранних стадиях, когда пациенты еще не проявляют сильных симптомов, что может привести к задержки диагностики и лечения заболевания. Кроме того, эти методы являются дорогостоящими и могут быть не доступны для большинства людей. В связи с этим разработка программы для диагностики БП средствами машинного обучения является важной задачей в медицинской практике.

## Набор данных HandPD

Field study: Unesp 2010  
University of Applied Sciences  
Regensburg  
Biometric Smart Pen Project  
Universidade Estadual Paulista  
Faculdade de Medicina (FMB),  
Botucatu

No: \_\_\_\_\_  
RG: \_\_\_\_\_

Idade: \_\_\_\_\_ Mão dominante: ( ) direita ( ) esquerda

Desenhar círculo 12 vezes no mesmo lugar sem parar.	Desenhar círculo no ar 12 vezes no mesmo lugar sem parar.
Desenhar espiral após sinal sonoro, de dentro para fora.	
Desenhar meander após sinal sonoro, de dentro para fora.	
Diadococinesia: Mão direita 20 segundos.	
Diadococinesia: Mão esquerda 20 segundos.	

Пустой бланк с шаблонами

22/11/2010

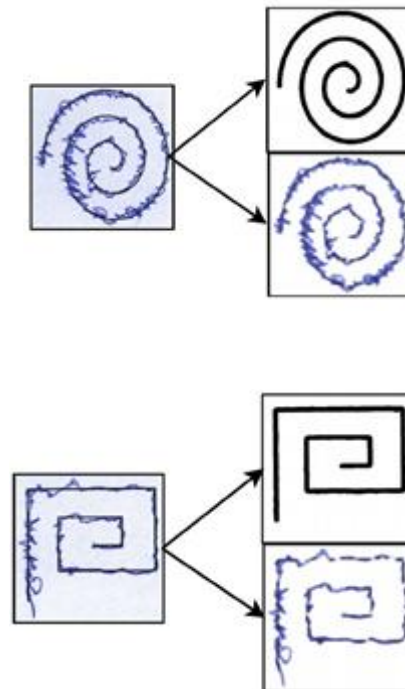
Field study: Unesp 2010  
University of Applied Sciences  
Regensburg  
Biometric Smart Pen Project  
Universidade Estadual Paulista  
Faculdade de Medicina (FMB),  
Botucatu

No: \_\_\_\_\_  
RG: \_\_\_\_\_

Idade: 56 Mão dominante: ☒ direita ( ) esquerda

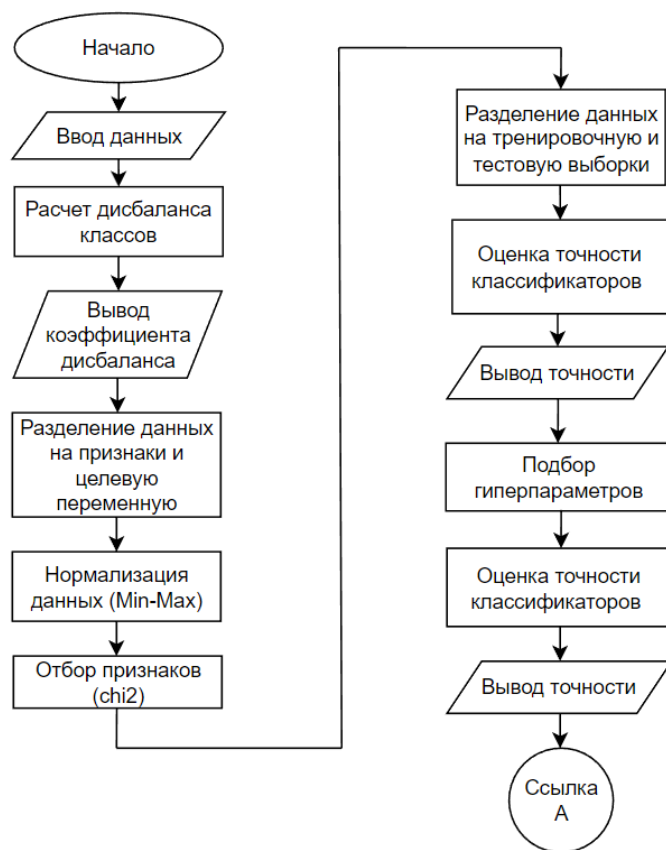
Desenhar círculo 12 vezes no mesmo lugar sem parar.	Desenhar círculo no ar 12 vezes no mesmo lugar sem parar.
Desenhar espiral após sinal sonoro, de dentro para fora.	
Desenhar meander após sinal sonoro, de dentro para fora.	
Diadococinesia: Mão direita 20 segundos.	
Diadococinesia: Mão esquerda 20 segundos.	

Бланк заполненный  
пациентом с БП



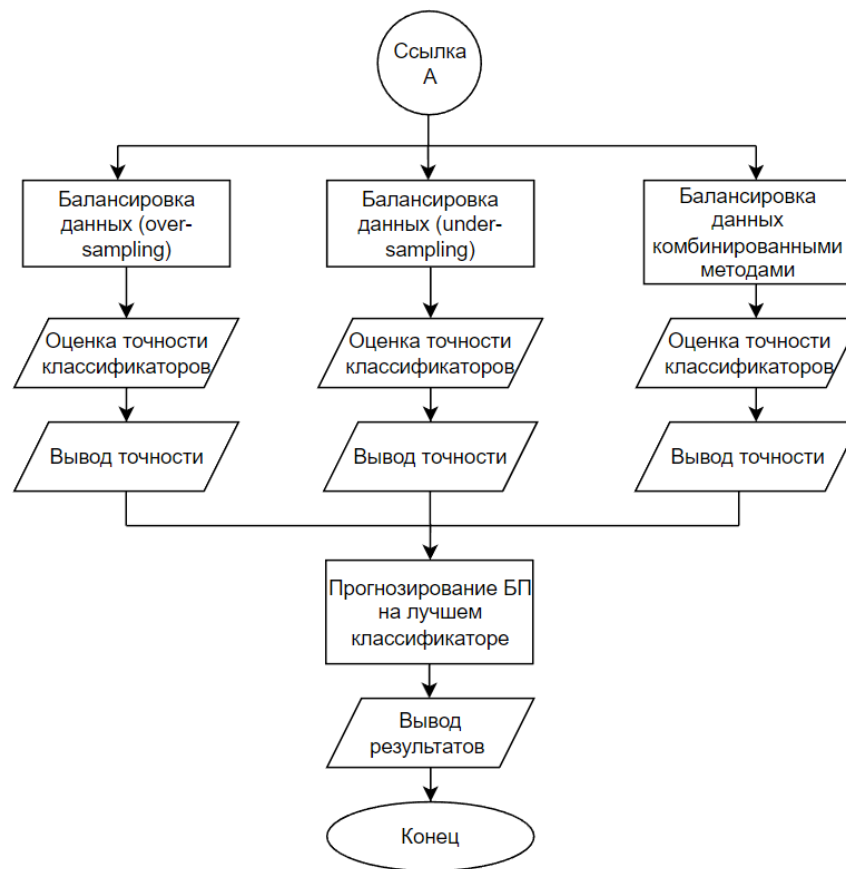
Результаты  
извлечения  
шаблонного и  
рукописного следа

## Блок-схема разработанной программы для диагностики БП



Блок схема программы

## Блок-схема разработанной программы для диагностики БП

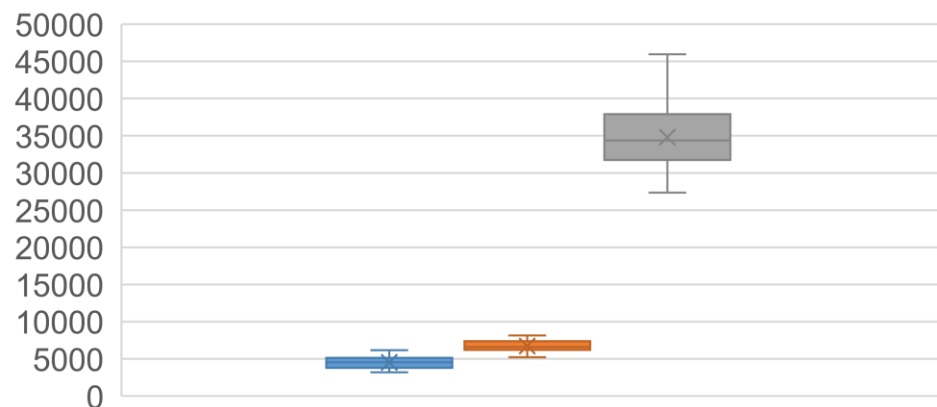


Блок схема программы

# Нормализация данных

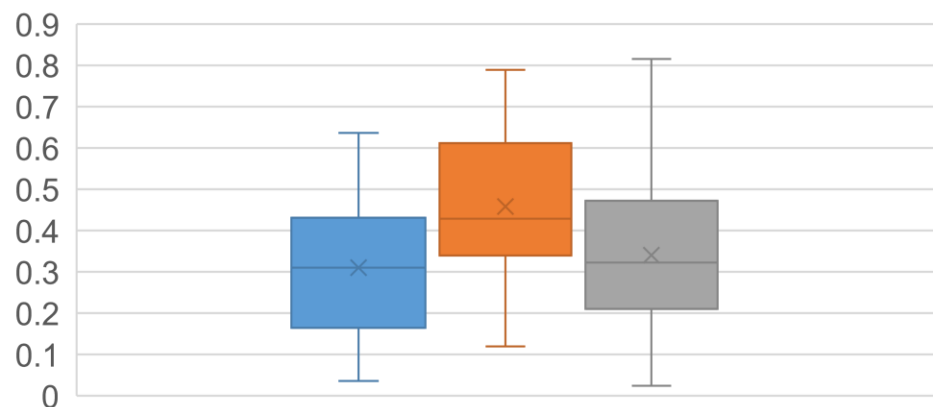
Данные до нормализации

■ a0 ■ a1 ■ a2



Данные после нормализации

■ a0 ■ a1 ■ a2



Методы нормализации данных:

- Min-Max нормализация
- Z-score нормализация

## Отбор признаков

Отбор признаков - это процесс выбора наиболее значимых признаков из набора данных для использования в модели машинного обучения. Этот процесс позволяет улучшить качество модели, сократить время обучения и снизить риск переобучения.

```
def show_best_k_value(
    filtering_alg: Any,
    x: pd.DataFrame,
    y: pd.Series,
    classifiers: list[Any]
) -> pd.DataFrame:

    y_copy = y.copy(deep=True)
    x_copy = x.copy(deep=True)

    max_balanced_accuracy = {}
    best_parameters_count = {}
    for clf in classifiers:
        max_balanced_accuracy[clf] = 0
        best_parameters_count[clf] = 0

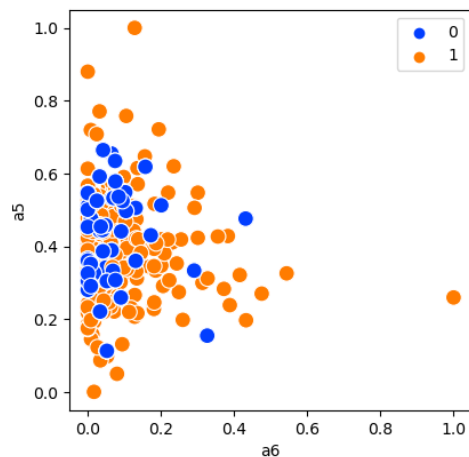
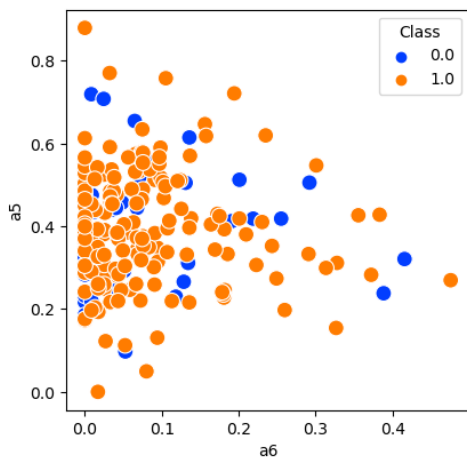
    features_after_selection = None
    for k in np.arange(1, x.shape[1]):
        print(f"Parameters count {k}")
        features_after_selection = get_features_after_selection(filtering_alg, k, x_copy, y_copy)
        for clf in classifiers:
            balanced_accuracy = cross_validating_score(clf, features_after_selection, y_copy)[0]
            if max_balanced_accuracy[clf] < balanced_accuracy:
                max_balanced_accuracy[clf] = balanced_accuracy
                best_parameters_count[clf] = k
        print('\n')

    for key in max_balanced_accuracy.keys():
        print(f'Parameters count {best_parameters_count[key]} \n Classifier: {key} \n value = {max_balanced_accuracy[key]}'
```

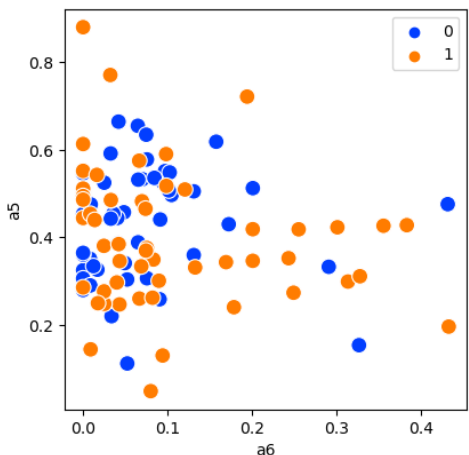
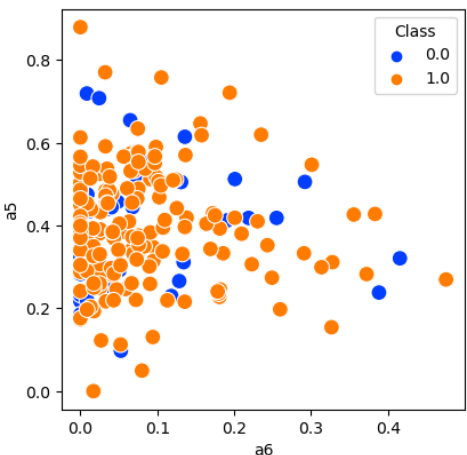
Код отвечающий за отбор признаков



## Балансировка данных



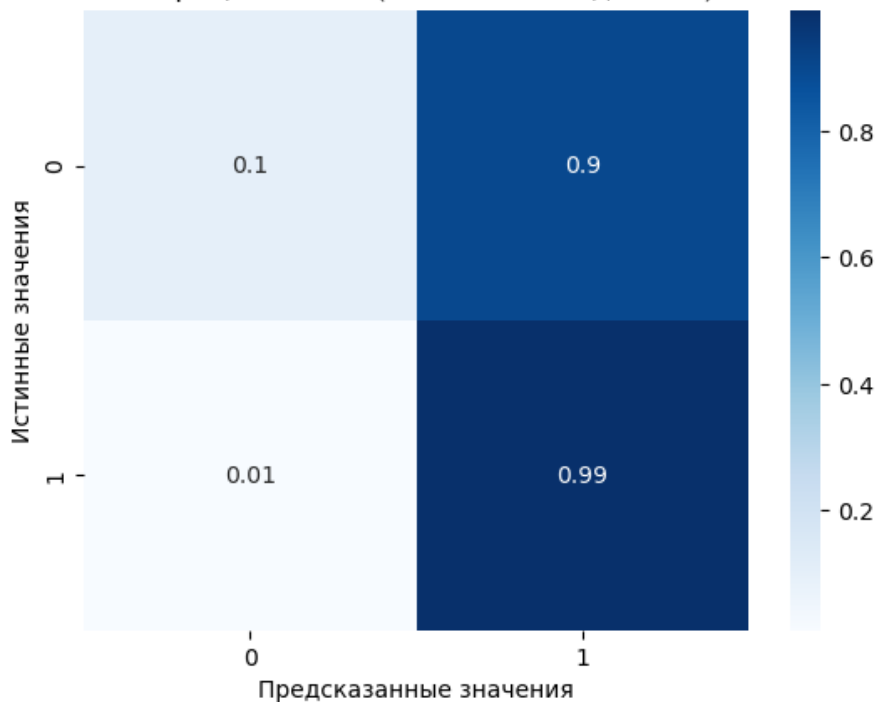
Пример работы метода добавления данных (over-sampling)



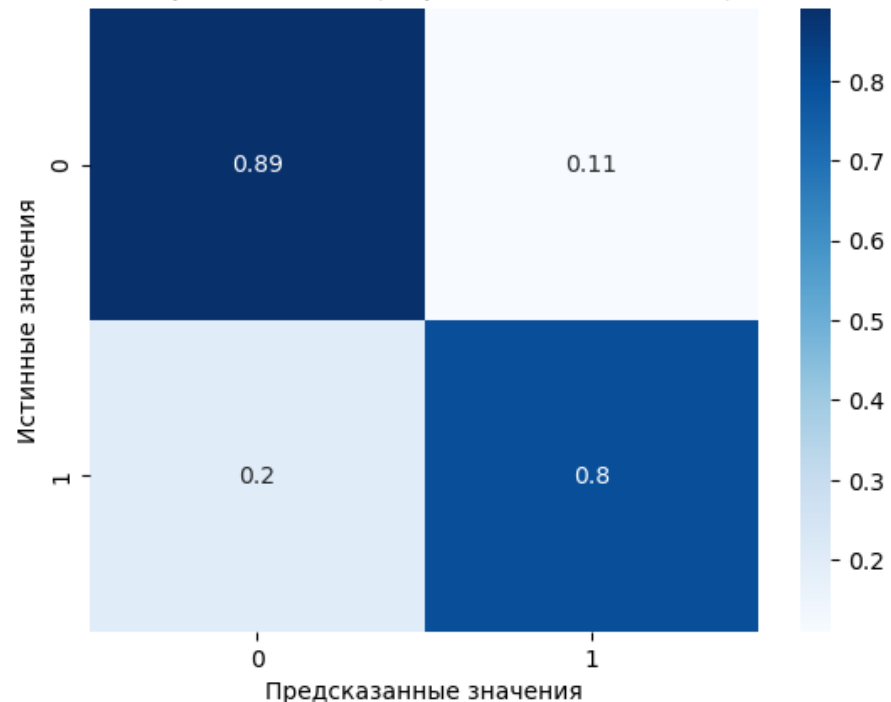
Пример работы метода удаления данных (under-sampling)

## Оценка точности

Матрица ошибок (Изначальные данные)



Матрица ошибок (Обработанные данные)



## Результаты экспериментов

<i>Classifier</i>	<i>BA, %</i>	<i>Accuracy, %</i>	<i>F1, %</i>	<i>FP, %</i>	<i>FN, %</i>
LogisticRegression	90,02	<b>92,35</b>	87,05	<b>4,05</b>	15,56
DecisionTreeClassifier	84,08	88,45	78,72	4,05	27,78
MLPClassifier	89,14	91,31	85,42	5,05	16,67
MultinomialNB	84,09	86,02	77,58	10,71	21,11
SVC	<b>91,75</b>	91,98	<b>87,72</b>	7,69	<b>8,89</b>

## Результаты полученные с помощью разработанной программы

<i>Classifier</i>	<i>Control group, Accuracy %</i>	<i>Patient group, Accuracy %</i>	<i>Global, Accuracy %</i>
<i>OPF</i>	64,96 ± 16,27	60,23 ± 4,73	<b>55,86 ± 3,63</b>
<i>NB</i>	27,30 ± 37,36	70,36 ± 39,08	<b>45,79 ± 4,15</b>
<i>SVM,</i>	12,50 ± 25,00	96,49 ± 2,50	<b>58,61 ± 2,84</b>

<i>Classifier</i>	<i>Accuracy, %</i>
ГП	72,36
AdaBoost + chisquare	76,44
CNN	<b>79,62</b>
LR, SVM	Меандры: 72,16 Спираль: 77,45

### Результаты из исследования (А)

### Результаты из исследования (Б)

## Пример прогнозов выдаваемых программой

	A	B
1	<b>0</b>	0
2	<b>1</b>	1
3	<b>2</b>	1
4	<b>3</b>	1
5	<b>4</b>	1
6	<b>5</b>	1
7	<b>6</b>	1
8	<b>7</b>	1
9	<b>8</b>	1
10	<b>9</b>	1
11	<b>10</b>	1
12	<b>11</b>	1
13	<b>12</b>	1
14	<b>13</b>	0
15	<b>14</b>	1
16	<b>15</b>	1
17	<b>16</b>	1

Результаты прогнозов

```

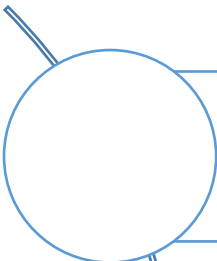
1 features_before_sampling = features_train.copy(deep=True)
2 target_before_sampling = target_train.copy(deep=True)
3 features_after_sampling, target_after_sampling = resample_data(
4     SMOTEENN(sampling_strategy='auto', random_state=random_state),
5     "SMOTEENN",
6     features_before_sampling,
7     target_before_sampling,
8 )
9
10 svc.fit(features_after_sampling, target_after_sampling)
11 target_predict = svc.predict(features_test)
12
13 with pd.ExcelWriter('predictions.xlsx') as writer:
14     pd.Series(target_predict).to_excel(writer, sheet_name='Спрогнозированные данные')

```

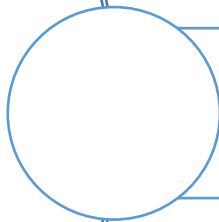
Участок кода отвечающий за выдачу прогнозов

## Заключение

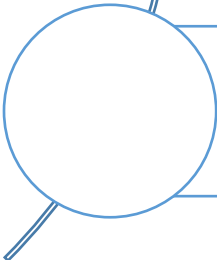
Разработанная программа позволяет:



**Увеличить точность** предсказаний пациентов с БП минимум на **12,39%** по сравнению с аналогами.



**Снизить стоимость диагностики БП** и тем самым повысить доступность для большинства людей.



**Внести вклад в будущие исследования.** Так как исследователи могут продолжить совершенствование и развитие этой технологии, для повышения точности и доступности диагностики.

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И  
РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра компьютерных систем в управлении и проектировании (КСУП)

**СПАСИБО ЗА ВНИМАНИЕ!**

Выполнил: студент 4 курса, группы 589-3  
Пахомов Максим Владимирович

Руководитель работы: доцент кафедры КСУП, к.т.н.  
Бардамова Марина Борисовна

Томск 2023

## Признаки в наборе данных (1)

Признак  $a_0$  – Среднеквадратичное отклонение (СКО) разницы между рукописный следом (НТ) и радиусом экзаменационного шаблона (ЕТ):

$$a_0 = \sqrt{\frac{1}{n} \sum_{i=0}^n (r_{HT}^i - r_{ST}^i)^2}$$

где  $n$  – количество случайно выбранных точек на линии рисунка,

$i$  – точка на рукописной и шаблонной линиях,

$r_{HT}^i$  – расстояние (радиус) от центра рисунка и до точки  $i$  на рукописном рисунке,

$r_{ST}^i$  – расстояние от точки  $i$  на шаблонной линии до центра рисунка.

## Признаки в наборе данных (2)

Признаки  $a_1$  и  $a_2$  – максимальное и минимальное различие между радиусами  $r_{HT}^i$  и  $r_{ST}^i$ .

$$a_1 = \max(\{|r_{HT}^1 - r_{ST}^1|, \dots, |r_{HT}^n - r_{ST}^n|\}), a_2 = \min(\{|r_{HT}^1 - r_{ST}^1|, \dots, |r_{HT}^n - r_{ST}^n|\})$$

Признак  $a_3$  – стандартное отклонение различий между радиусами  $r_{HT}^i$  и  $r_{ST}^i$ .

Признак  $a_4$  – Средний относительный тремор (MRT), определяемый как средняя разница между радиусом данного образца и его  $d$  ближайших соседей слева.

$$\bullet \quad a_4 = \frac{1}{n-d} \sum_{i=d}^n |r_{HT}^i - r_{HT}^{i-d+1}|$$

Признаки  $a_5$  и  $a_6$  – максимальное и минимальное значение радиуса НТ,  $a_7$  – стандартное отклонение значений НТ.

Признак  $a_8$  – количество раз, когда разница между НТ и ЕТ радиусом меняется с отрицательного на положительный, или наоборот.



## Точность экспериментов до балансировки

Результаты экспериментов без балансировки и подбора гиперпараметров

<i>Classifier</i>	<i>BA, %</i>	<i>Accuracy, %</i>	<i>F1, %</i>	<i>FP, %</i>	<i>FN, %</i>
LogisticRegression	53,83	81,92	89,9	92,33	<b>0</b>
DecisionTreeClassifier	<b>61,18</b>	76,44	85,41	<b>63,67</b>	13,97
MLPClassifier	57,65	82,59	90,1	83,33	1,36
MultinomialNB	50	80,45	89,16	100	<b>0</b>
SVC	56,27	<b>82,63</b>	<b>90,23</b>	87	0,45

Результаты экспериментов с подобранными гиперпараметрами, но без балансировки

<i>Classifier</i>	<i>BA, %</i>	<i>Accuracy, %</i>	<i>F1, %</i>	<i>FP, %</i>	<i>FN, %</i>
LogisticRegression	67,85	84,38	<b>90,71</b>	59,33	<b>4,96</b>
DecisionTreeClassifier	68,53	77,58	85,63	46,33	16,6
MLPClassifier	70,17	<b>84,39</b>	90,53	53,33	6,32
MultinomialNB	68,37	73,9	82,43	40,67	22,59
SVC	<b>72,77</b>	78,94	86,18	<b>37,33</b>	17,33

## Балансировщики

### Добавление экземпляров (Over-sampling)

- RandomOverSampler, SMOTE, ADASYN

### Удаление экземпляров (Under-sampling)

- RandomUnderSampler, CondensedNearestNeighbour, TomekLinks, ClusterCentroids

### Комбинирование подходов:

- SMOTEENN, SMOTETomek

## Формулы для оценки точности

$$BA = \frac{TP_{rate} + TN_{rate}}{2}$$

$$\text{Accuracy (Точность)} = \frac{TP}{TP + TN + FP + FN}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Precision \text{ (Точность)} = \frac{TP_{rate}}{TP_{rate} + FP_{rate}}$$

$$Recall \text{ (Полнота)} = \frac{TP_{rate}}{TP_{rate} + FN_{rate}}$$

## Методы отбора признаков

Хи-квадрат (chi2) -  $\chi^2 = \sum \frac{(O-E)^2}{E}$ ,

Где О – наблюдаемая частота,

Е – ожидаемая частота

Взаимная информация (mutual\_info\_classif)

Статистический тест F-оценки (f\_classif)

## Нормализация

$$MinMax = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$Z_{score} = \frac{x - \mu}{\sigma},$$

Где  $x$  – исходное значение признака,

$\mu$  – среднее значение признака,

$\sigma$  – стандартное отклонение признака