**Лабораторная работа 3**

Бинарная и многоклассовая классификация. Оценка качества задачи классификации

Цель: «Изучение на практике методов data mining для решения задач классификации и оценке качества классификации»

**Содержание**

[Краткий опрос по теоретической часть 2](#_Toc188863171)

[Машинное обучение с учителем 2](#_Toc188863172)

[Задание на лабораторную работу 4](#_Toc188863173)

[Задача 1. Разбор данных из Kaggle 4](#_Toc188863174)

[Задача 1. Данные 5](#_Toc188863175)

[Варианты датасетов для задачи классификации 5](#_Toc188863176)

[Самостоятельная работа. Задача 2 6](#_Toc188863177)

[Варианты датасетов для замены 6](#_Toc188863178)

[Сложность выполнения 6](#_Toc188863179)

[Полезные ссылки 7](#_Toc188863180)

[Требования к отчёту по работе 7](#_Toc188863181)

# **Краткий опрос по теоретической часть**

В данном разделе необходимо ответить на вопросы.

1. Что такое классификация?
2. Чем отличается бинарная классификация от многоклассовой?
3. Какие преимущества имеет логистическая регрессия?
4. Какую задачу решает логистическая регрессия?
5. Как исключить входные факторы (признаки) с низкой значимостью?
6. Как оценить точность классификации?
7. Как работает алгоритм KNN? Чем отличается алгоритм от логистической регрессии?
8. Что показывают критерии качества и ROC-кривая для задачи с определением пола?
9. Какие входные и выходные параметры используются в модели Логистической регрессии?
10. Что показывают критерии качества и ROC-кривая для задачи с определением доброкачественной опухоли?
11. Как построить таблицу сопряженности (confusion matrix)?

# **Машинное обучение с учителем**

Обучение с учителем подразумевает предварительную разметку человеком объектов из обучающей выборки, т.е. присваивание этим объектам конечного числа меток для задач классификации или значений зависимой переменной для задач регрессии. Тогда задача классификации представляется как отнесение объектов тестовой выборки к одной из ранее присвоенных меток, т.е. к одному из классов.

Под задачей регрессии подразумевают определение математической модели (функции регрессии), устанавливающей функциональную связь между зависимой переменной и группой независимых переменных с учетом ошибки модели.

Все тексты программ, реализующих методы классификации и регрессии, для данной лабораторной работы можно скачать по url в разделе Полезные ссылки.

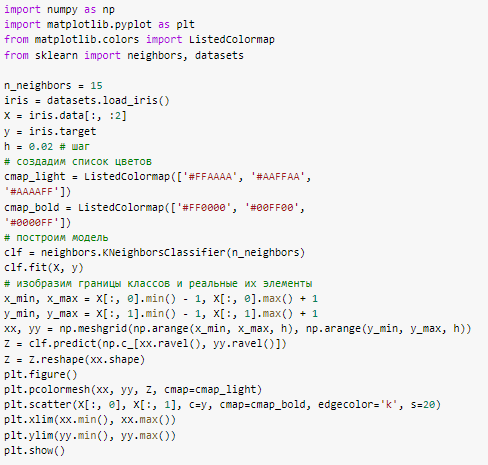


Рис. 1. Листинг настройки методы KNN для классификации в sklearn

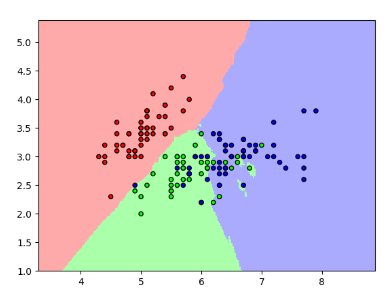


Рис. 2. Результат работы программы

(фон – границы классов, точки – элементы классов)

# **Задание на лабораторную работу**

## **Задача 1. Разбор данных из Kaggle**

Построить минимум две модели классификации с использованием алгоритмов KNN, логистическая регрессия, SVM. Алгоритмы допустимо использовать из библиотеки scikit-learn.

Модели строятся для задачи бинарной классификации.

Затем сравните их качество, используя метрики, такие как A, P, R, E, confusion\_matrix, ROC\_AUC.

на тестовых данных.

Подробное описание:

1. Загрузить один из наборов данных. Загружаются все файлы по ссылке.
2. Провести разведочный анализ данных, ответив на следующие вопросы:
3. Сколько строк в датафрейме, сколько столбцов
4. Сколько места занимает датафрейм в оперативной памяти
5. Для каждой интервальной переменной подсчитать следующее - мин, медиана, среднее, макс и персентили 25, 75
6. Для каждой категориальной переменной рассчитать моду и сколько раз мода встречается в данных
7. Подготовка датасета к построению моделей ML
8. Провести анализ и обработку пропусков (либо заменить, либо удалить)
9. Провести анализ и обработку выбросов (либо заменить, либо удалить)
10. Провести анализ и обработку категориальных переменных (сколько таких переменных, закодируйте категориальные переменные одним из методов ( one hot encoding, mean target, frequence encoding)
11. Разделить датасет на трейн и тест
12. Построить следующие классификационные алгоритмы:
13. Knn
14. Logistic regression
15. SVM
16. Оценить качество алгоритмов, выбрать самый оптимальный алгоритм
17. Возможно ли улучшить алгоритм, предложите идеи?

## **Задача 1. Данные**

Таблица 1.

### Варианты датасетов для задачи классификации

|  |  |
| --- | --- |
| Варианты | Ссылка |
| 1, 5, 9, 13, 17, 21 | Набор данных <https://www.kaggle.com/competitions/playground-series-s3e12/data?select=train.csv> |
| 2, 6, 10, 14, 18, 22 | Набор данных Титаник, можете скачать по -разному, вот один из вариантов По данным о пассажирах необходимо построить модель, которая спрогнозирует кто выжил, а кто нет  import seaborn as sns  df = sns.load\_dataset('titanic')  X = df.drop(columns='survived')  y = df["survived"] |
| 3, 7, 11, 15, 19, 23 | Набор данных <https://www.kaggle.com/competitions/playground-series-s3e10/data?select=train.csv> |
| 4, 8, 12, 16, 20, 24 | встроенный наборов данных, датасет об информации о раке груди. Необходимо по независимым признакам спрогнозировать есть рак или нет  from sklearn import datasets  data = datasets.load\_breast\_cancer()  X = pd.DataFrame(data["data"], columns=data["feature\_names"])  y = data["target"] |

## **Самостоятельная работа. Задача 2**

В качестве индивидуального набора данных взять данных из Лабораторной работы 1. Выполнить те же действия, что в Задаче 1.

Если по какой-то причине данные взять для анализа нельзя, то выбрать набора данных из представленных ниже. Для данных ниже описать какие есть столбцы, какие из них признаки и какой целевая переменная перед исследованием.

Таблица 2.

### Варианты датасетов для замены

|  |  |  |
| --- | --- | --- |
| № | База данных 2 | Оценка |
| 1 | Вино | <https://www.kaggle.com/datasets/yasserh/wine-quality-dataset> |
| 2 | Ирисы Фишера | <https://www.kaggle.com/datasets/uciml/iris> |
| 3 | Диабет | <https://www.kaggle.com/code/kunalvashisht/diabetes-visualization-with-seaborn/input> |

## **Сложность выполнения**

**Сложность: Rare**

* Реализовать только общую часть задания
* Построить 2 метода из 3

**Сложность: Medium**

* Реализовать общую часть задания
* Добавить нормализацию исходных данных для KNN и другому алгоритму, если понадобится
* Реализовать хотя бы 1 алгоритм для самостоятельной части задания по Лабораторной 1 или выбранному набору данных

**Сложность:** **Well-done**

* Реализовать общую и самостоятельную часть задания
* Добавить нормализацию исходных данных
* Реализованы 3 алгоритма и подробно описано исследование 2 датасетов.

# **Полезные ссылки**

1. Примеры работы с алгоритмами из лабораторной работы 2. URL: URL: <https://colab.research.google.com/drive/13I1PPFm-S8w40Mn-7CalLIaWcRoBjYCX#scrollTo=K8hOMjE4ZUko>
2. Тестовые данные к учебнику по анализу данных Себастьяна Рашки (Титаник, вина, ирисы Фишера). URL: <https://github.com/rasbt/pattern_classification/tree/master/data>
3. Тестовые датасеты из библиотеки sklearn (Бостон, Ирисы Фишера, вина, диабет, рак груди), . URL: <https://scikit-learn.ru/7-1-toy-datasets/>
4. Построение матрицы ошибок (confusion matrix). URL: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>

# **Требования к отчёту по работе**

1. Титульный лист;
2. Вариант задания;
3. База данных 1
   1. Описание базы данных – признаки, целевая переменная;
   2. Исследовательский анализ данных;
   3. Выводы по анализу;
   4. Описание алгоритмов;
   5. Визуализация результата;
   6. Критерии качества;
   7. Ответы на контрольные вопросы;
   8. Вывод по результату и критериям.
4. База данных 2
   1. Описание базы данных – признаки, целевая переменная;
   2. Исследовательский анализ данных;
   3. Выводы по анализу;
   4. Описание алгоритмов;
   5. Визуализация результата;
   6. Критерии качества;
   7. Ответы на контрольные вопросы;
   8. Вывод по результату и критериям.