Министерство науки и образования РФ

Федеральное государственное бюджетное учреждение

высшего образования

**«Тверской государственный технический университет»**

(ТвГТУ)

Кафедра программного обеспечения

**Отчет по лабораторной работе №3**

По дисциплине: «Анализ больших данных»

Тема: «Бинарная и многоклассовая классификация. Оценка качества задачи классификации»

|  |
| --- |
| Выполнила:  студентка группы  Б.ПИН.РИС - 21.06  Александрова М. И. |
| Проверила:  старший преподаватель  кафедры ПО  Корнеева Е.И. |

Тверь 2024

**Оглавление**

[**Ссылка на программу** 3](#_Toc194254347)

[**Теоретические вопросы** 3](#_Toc194254348)

[**Задача 1. Разбор данных из Kaggle** 7](#_Toc194254349)

[**Разведочный анализ данных** 8](#_Toc194254350)

[**Построение классификационных моделей** 9](#_Toc194254351)

[**Самостоятельная работа. Задача 2** 10](#_Toc194254352)

[**Разведочный анализ данных** 10](#_Toc194254353)

[**Построение классификационных моделей** 11](#_Toc194254354)

# **Ссылка на программу**

<https://github.com/milana-cat/BigData.git>

# **Теоретические вопросы**

1. **Что такое классификация?**

Классификация — это задача машинного обучения, при которой объекты разделяются на заранее определенные категории (классы) на основе их признаков.

1. **Чем отличается бинарная классификация от многоклассовой?**

Бинарная классификация: два возможных класса (например, «да» или «нет», «болен» или «здоров»). Многоклассовая классификация: три и более классов (например, определение породы животного или типа транспортного средства).

1. **Какие преимущества имеет логистическая регрессия?**

* Простота интерпретации
* Быстрота вычислений
* Хорошо работает с линейно разделимыми данными
* Выдает вероятности принадлежности к классу

1. **Какую задачу решает логистическая регрессия?**

Логистическая регрессия решает задачу классификации, предсказывая вероятность принадлежности объекта к одному из классов.

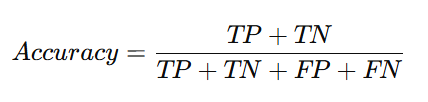
1. **Как исключить входные факторы (признаки) с низкой значимостью?**

* Анализ коэффициентов модели
* Использование регуляризации (L1, Lasso)
* Методы отбора признаков (например, анализ значимости признаков)

1. **Как оценить точность классификации?**

Для оценки качества классификации используются несколько метрик. Выбор конкретной метрики зависит от задачи и важности разных типов ошибок.

1. Accuracy (общая точность)



Где:

TP (True Positive) — верно предсказанные положительные примеры.

TN (True Negative) — верно предсказанные отрицательные примеры.

FP (False Positive) — ложноположительные предсказания (ошибочное определение объекта как положительного).

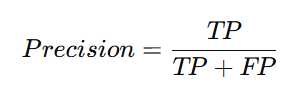
FN (False Negative) — ложноотрицательные предсказания (ошибочное определение объекта как отрицательного).

Когда использовать: Если классы сбалансированы (примерно равное количество объектов в каждом классе).

Когда не использовать: Если один класс встречается гораздо чаще другого (например, при диагностике редких заболеваний).

2. Precision (точность положительного класса)

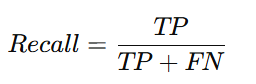
Показывает, насколько точны предсказания положительного класса.



Когда использовать: Если ложные срабатывания нежелательны (например, при выдаче рекламы: лучше показать меньше, но более релевантных объявлений).

3. Recall (полнота)

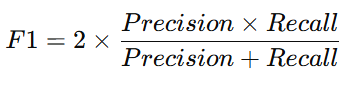
Показывает, насколько хорошо модель находит все положительные примеры.



Когда использовать: Если важно находить все объекты положительного класса (например, при диагностике рака — лучше ошибиться с ложноположительным результатом, чем пропустить болезнь).

4. F1-score (сбалансированная метрика)

F1-score используется, когда важно учитывать и precision, и recall.



Когда использовать: Если классы несбалансированы.

5. ROC-кривая и AUC (площадь под кривой ROC)

ROC-кривая (Receiver Operating Characteristic) показывает, как меняется полнота (recall) в зависимости от доли ложных срабатываний (False Positive Rate).

AUC (Area Under Curve) — площадь под ROC-кривой, показывает, насколько хорошо модель отделяет один класс от другого.

AUC = 0.5 → модель случайная.

AUC = 1 → идеальная модель.

Когда использовать: Если важно оценить качество вероятностных предсказаний модели.

1. **Как работает алгоритм KNN? Чем отличается алгоритм от логистической регрессии?**

KNN (k ближайших соседей): новый объект классифицируется по классу большинства среди его k ближайших соседей.

KNN — метод на основе расстояний, не требует обучения; логистическая регрессия — параметрический метод, строит функцию разделения.

1. **Что показывают критерии качества и ROC-кривая для задачи с определением пола?**

Аccuracy, precision, recall показывают, насколько хорошо модель различает мужчин и женщин. ROC-кривая показывает баланс между чувствительностью (recall) и специфичностью при разных порогах.

1. **Какие входные и выходные параметры используются в модели логистической регрессии?**

Входные параметры: числовые или категориальные признаки объекта.

Выходные параметры: вероятность принадлежности к определенному классу.

1. **Что показывают критерии качества и ROC-кривая для задачи с определением доброкачественной опухоли?**

Accuracy, precision, recall показывают, насколько точно модель определяет доброкачественные и злокачественные опухоли.

ROC-кривая позволяет выбрать оптимальный порог классификации.

1. **Как построить таблицу сопряженности (confusion matrix)?**
2. Определить количество истинно положительных (TP), истинно отрицательных (TN), ложно положительных (FP) и ложно отрицательных (FN) классификаций.
3. Оформить их в таблицу.

## **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.

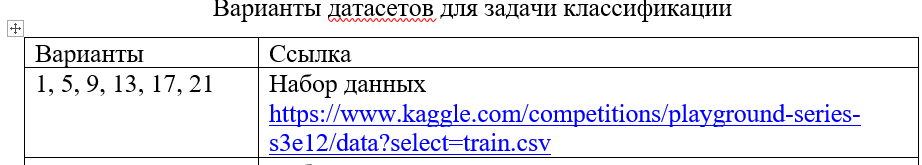
Модели строятся для задачи бинарной классификации.

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

на тестовых данных.

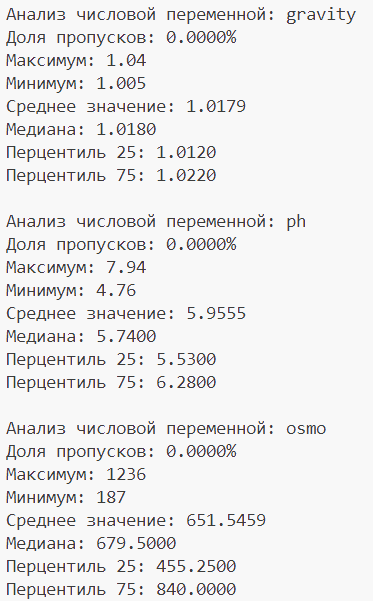
Подробное описание:

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?

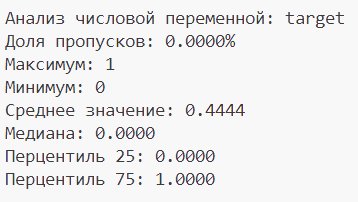


# **Разведочный анализ данных**

1. Сколько строк в датафрейме, сколько столбцов
2. Сколько места занимает датафрейм в оперативной памяти
3. Для каждой интервальной переменной подсчитать следующее - мин, медиана, среднее, макс и персентили 25, 75







1. Для каждой категориальной переменной рассчитать моду и сколько раз мода встречается в данных

Категориальных переменных в текущем датасете не было.

# **Построение классификационных моделей**

1. Knn

Оптимальное значение k: 20

Confusion Matrix:

[[27 18]

[14 24]]

Accuracy: 0.6145

Precision: 0.6145

Recall: 0.6145

F1 Score: 0.6145

ROC-AUC: 0.6257

1. Logistic regression

Confusion Matrix:

[[34 11]

[24 14]]

Accuracy: 0.5783

Precision: 0.5783

Recall: 0.5783

F1 Score: 0.5783

ROC-AUC: 0.6053

1. SVM

Confusion Matrix:

[[45 0]

[38 0]]

Accuracy: 0.5422

Precision: 0.5422

Recall: 0.5422

F1 Score: 0.5422

Вывод:

KNN — лучший алгоритм в данной задаче и показывает наивысшие метрики. Логистическая регрессия немного уступает, но тоже показывает приемлемые результаты. SVM явно не подходит, так как полностью игнорирует один из классов (проблема линейно неразделимых данных).

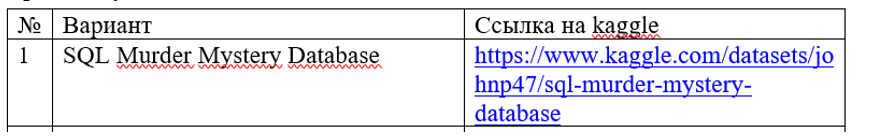
В качестве улучшения SVM, можно попробовать использовать другие

## **Самостоятельная работа. Задача 2**

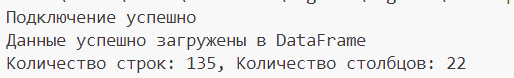
В качестве индивидуального набора данных взять данных из Лабораторной работы 1. Выполнить те же действия, что в Задаче 1.

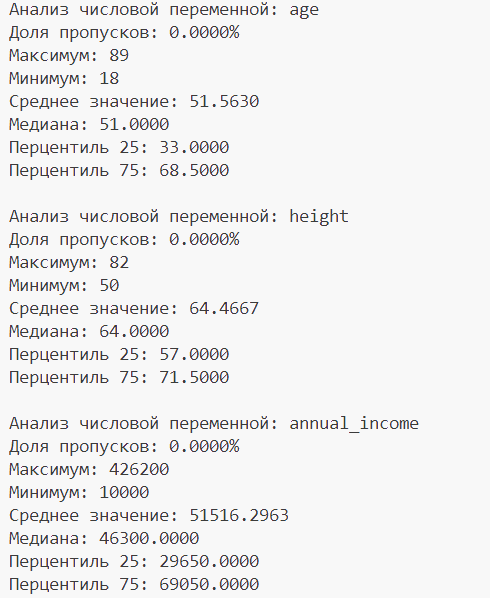
Если по какой-то причине данные взять для анализа нельзя, то выбрать набора данных из представленных ниже. Для данных ниже описать какие есть столбцы, какие из них признаки и какой целевая переменная перед исследованием.

Вариант 1

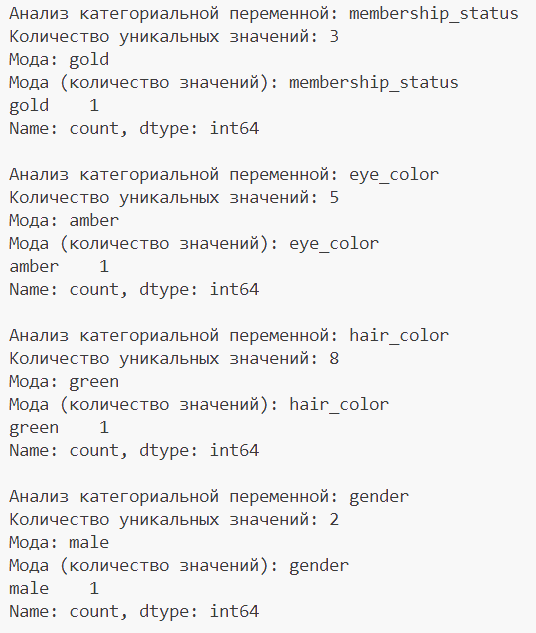


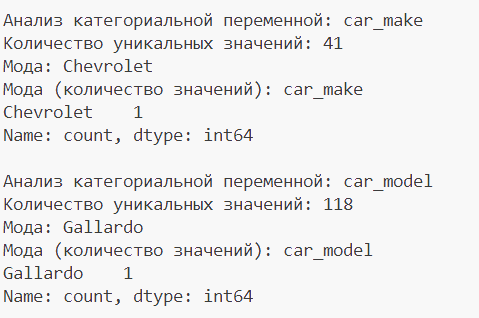
# **Разведочный анализ данных**

1. Сколько строк в датафрейме, сколько столбцов 
2. Сколько места занимает датафрейм в оперативной памяти 
3. Для каждой интервальной переменной подсчитать следующее - мин, медиана, среднее, макс и персентили 25, 75



1. Для каждой категориальной переменной рассчитать моду и сколько раз мода встречается в данных





# **Построение классификационных моделей**

1. Knn

Оптимальное значение k: 2

Confusion Matrix:

[[6 5 0]

[3 8 0]

[2 2 1]]

Accuracy: 0.5556

Precision: 0.6929

Recall: 0.4909

F1 Score: 0.4981

1. Logistic regression

Confusion Matrix:

[[ 0 11 0]

[ 3 8 0]

[ 2 2 1]]

Accuracy: 0.3333

Precision: 0.4603

Recall: 0.3091

F1 Score: 0.2778

1. SVM

Confusion Matrix:

[[1 7 3]

[0 9 2]

[1 4 0]]

Accuracy: 0.3704

Precision: 0.3167

Recall: 0.3030

F1 Score: 0.2448

Вывод:

KNN — лучший алгоритм в данной задаче и показывает наивысшие метрики. Логистическая регрессия и SVM работают плохо, вероятно, из-за линейной разделимости классов. Проблема с третьим классом – все алгоритмы плохо его классифицируют. В качестве улучшения SVM, можно попробовать использовать другие ядра. Для улучшения линейной регрессии, можно попробовать провести L1 и L2 регуляризации.