PETER THE GREAT
SAINT-PETERSBURG POLYTECHNIC UNIVERSITY

LDA
Glass Classification

Липс Екатерина Константиновна

гр. 5030102/10002

POLYTECH

Постановка задачи

Цель - классификация типов стекла на основе различных характеристик

Должно быть выполнено:
1. Исследовательский анализ данных (EDA)
2. Предварительная обработка данных
3. LDA

```
RI,Na,Mg,Al,Si,K,Ca,Ba,Fe,Type
1.52101,13.64,4.49,1.1,71.78,0.06,8.75,0,0,1
1.51761,13.89,3.6,1.36,72.73,0.48,7.83,0,0,1
1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0,0,1
1.51766,13.21,3.69,1.29,72.61,0.57,8.22,0,0,1
1.51742,13.27,3.62,1.24,73.08,0.55,8.07,0,0,1
1.51596,12.79,3.61,1.62,72.97,0.64,8.07,0,0.26,1
1.51743,13.3,3.6,1.14,73.09,0.58,8.17,0,0,1
1.51756,13.15,3.61,1.05,73.24,0.57,8.24,0,0,1
1.51918,14.04,3.58,1.37,72.08,0.56,8.3,0,0,1
1.51755,13,3.6,1.36,72.99,0.57,8.4,0,0.11,1
1.51571,12.72,3.46,1.56,73.2,0.67,8.09,0,0.24,1
1.51763,12.8,3.66,1.27,73.01,0.6,8.56,0,0,1
1.51589,12.88,3.43,1.4,73.28,0.69,8.05,0,0.24,1
```

data.csv

POLYTECH

Линейный дискриминантный анализ (LDA)

Линейный дискриминантный анализ (LDA) — алгоритм классификации и понижения размерности, позволяющий производить разделение классов наилучшим образом. Основная идея LDA заключается в предположении о многомерном нормальном распределении признаков внутри классов и поиске их линейного преобразования, которое максимизирует межклассовую дисперсию и минимизирует внутриклассовую.

1) для всех классов рассчитываются априорные вероятности и средние значения признаков;

2) на основе полученных значений рассчитываются (ковариационные) матрицы разброса между классами и внутри классов;

3) находятся собственные вектора и значения для линейного дискриминанта Фишера, который определяется отношением матриц из шага 2;

4) собственные вектора сортируются в порядке убывания в соответствии с собственными значениями и называются *дискриминантными векторами*, с помощью которых рассчитываются веса модели;

5) на основе полученных весов и априорных вероятностей рассчитывается вектор смещения;

6) новое пространство признаков меньшей размерности представляет из себя линейную комбинацию исходных признаков и дискриминантных векторов и называется *дискриминантным подпространством*;

7) спрогнозированные классы являются максимальной оценкой линейной комбинации тестовой выборки и весов + смещение.

Описание алгоритма

1. Исследовательский анализ данных (EDA):

- Отображение основных статистик и пропущенных значений.
- Построение графиков распределения типов стекла и корреляции между признаками.

2. Предварительная обработка данных:

- Масштабирование признаков.
- Разделение данных на обучающую и тестовую выборки.

3. Построение и оценка модели:

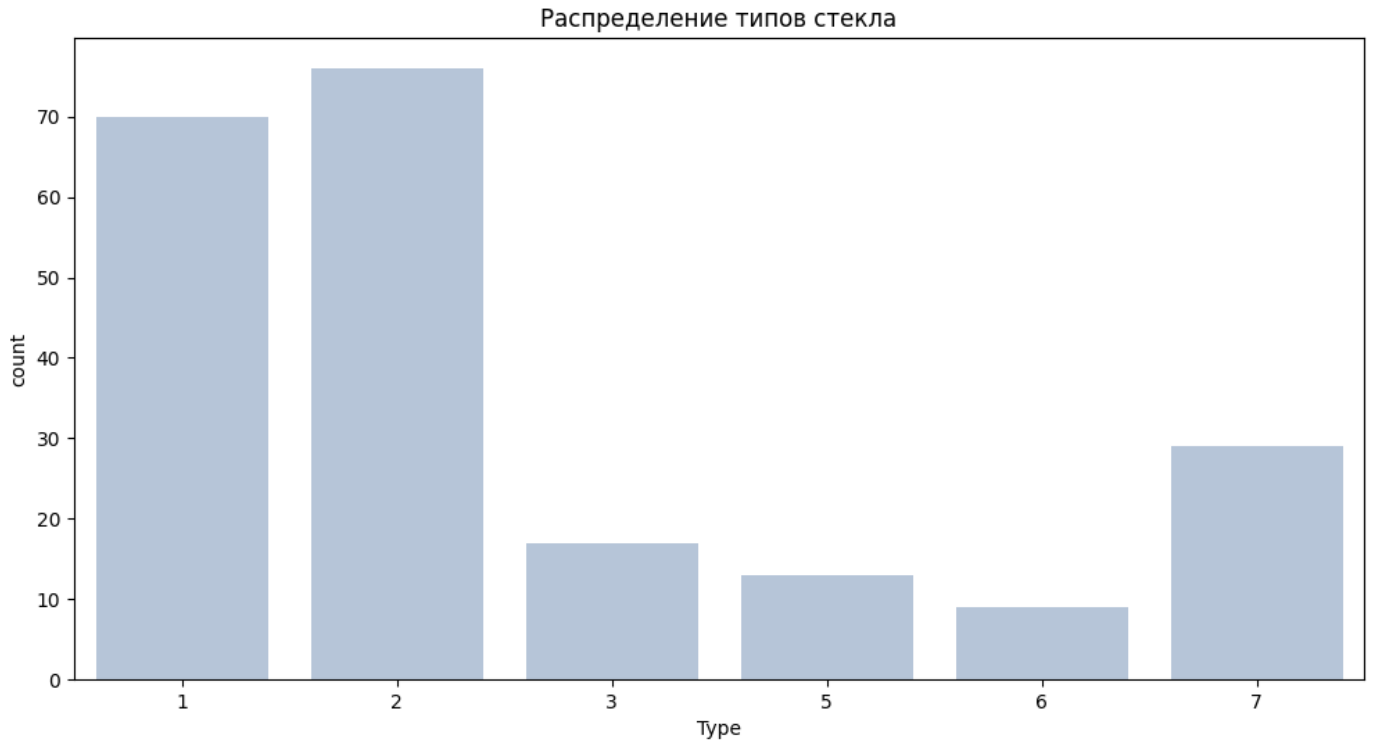
- Создание модели LDA.
- Оценка модели с использованием отчета классификации, матрицы ошибок и точности.

Исследовательский анализ данных (EDA)

Статистика данных:

	RI	Na	Mg	...	Ba	Fe	Type
count	214.000000	214.000000	214.000000	...	214.000000	214.000000	214.000000
mean	1.518365	13.407850	2.684533	...	0.175047	0.057009	2.780374
std	0.003037	0.816604	1.442408	...	0.497219	0.097439	2.103739
min	1.511150	10.730000	0.000000	...	0.000000	0.000000	1.000000
25%	1.516522	12.907500	2.115000	...	0.000000	0.000000	1.000000
50%	1.517680	13.300000	3.480000	...	0.000000	0.000000	2.000000
75%	1.519157	13.825000	3.600000	...	0.000000	0.100000	3.000000
max	1.533930	17.380000	4.490000	...	3.150000	0.510000	7.000000

Исследовательский анализ данных (EDA)

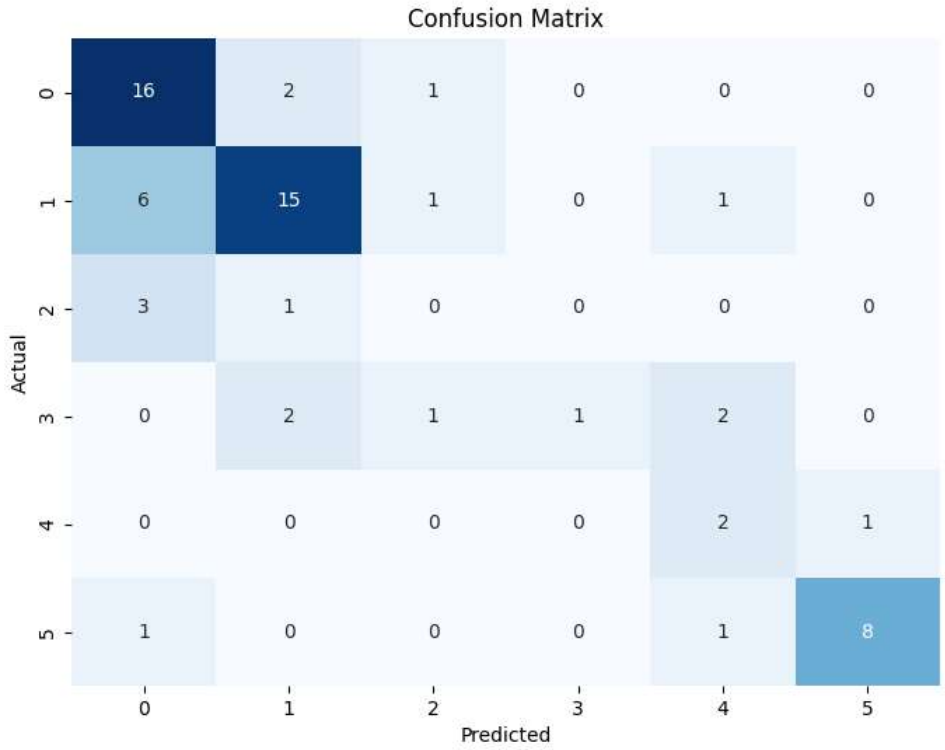


POLYTECH

Исследовательский анализ данных (EDA)

POLYTECH

Построение и оценка модели



Выводы:

- Типы стекла 1 и 2 классифицируются относительно хорошо, с высокой долей правильных предсказаний.
- Проблемы возникают с классификацией типов 3 и 5, где модель предсказывает мало или вообще не предсказывает эти классы правильно.
- Тип 7 классифицируется лучше всего (высокая точность и полнота).

POLYTECH

Построение и оценка модели

Classification Report:

	precision	recall	f1-score	support
1	0.62	0.84	0.71	19
2	0.75	0.65	0.70	23
3	0.00	0.00	0.00	4
5	1.00	0.17	0.29	6
6	0.33	0.67	0.44	3
7	0.89	0.80	0.84	10
accuracy			0.65	65
macro avg	0.60	0.52	0.50	65
weighted avg	0.69	0.65	0.63	65

Accuracy: 0.65

Выводы:

Precision и Recall для типов стекла сильно варьируются:

- Типы 1, 2, и 7 имеют сравнительно высокую precision и recall.
- Тип 3 вообще не классифицируется моделью, что указывает на серьезные проблемы с этим классом, возможно, из-за недостаточного количества данных.
- Тип 5 имеет высокую precision, но низкий recall, что говорит о том, что модель склонна игнорировать этот тип.

Возможные причины низкой производительности

- **Дисбаланс классов**:
  Некоторые типы стекла (например, 3, 5, 6) имеют мало примеров в данных, что затрудняет их классификацию.

- **Сложность данных**:
  Корреляция между признаками может быть сложной для модели LDA, которая предполагает линейную разделимость.

- **Шум в данных**:
  Возможны проблемы с качеством данных, например, пересечение признаков между классами.