Санкт-Петербургский политехнический университет
Петра Великого

Практическое задание по курсу
«Основы машинного обучения»

Классификация опухолей молочной железы: PCA и логистическая регрессия

Выполнил: Дмитриев Михаил
Группа: 5030102/10201

Санкт-Петербург
2024 г.

Введение и подготовка данных

Цель исследования:

- Классификация опухолей молочной железы на злокачественные и доброкачественные.

Используемые методы:

- Анализ главных компонент (PCA) — уменьшение размерности и визуализация данных.
- Логистическая регрессия — предсказание класса опухоли.

Предварительная обработка данных:

- Данные: *Breast Cancer Wisconsin Dataset* `https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data`.
- Удалены ненужные столбцы (*ID*, *Unnamed: 32*).
- Пропущенные значения заменены средними значениями.
- Целевая переменная преобразована: M (злокачественные) $= 1$, B (доброкачественные) $= 0$.
- Признаки нормализованы с использованием **StandardScaler**.

Методы: PCA и логистическая регрессия

Анализ главных компонент (PCA):

- Метод уменьшения размерности данных.
- Основные шаги:
  - Вычисление ковариационной матрицы данных.
  - Нахождение собственных векторов и собственных значений.
  - Проекция данных на пространство главных компонент.
- Цель: выделить главные компоненты, объясняющие максимальную дисперсию данных.

Логистическая регрессия:

- Метод классификации, основанный на логистической функции:

$$h(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}}$$

- Прогнозирует вероятность принадлежности объекта к одному из классов.
- Цель: минимизировать логарифмическую функцию потерь.

Анализ главных компонент (PCA)

- PCA позволяет уменьшить размерность данных и выделить основные компоненты.
- Данные преобразованы в двумерное пространство.

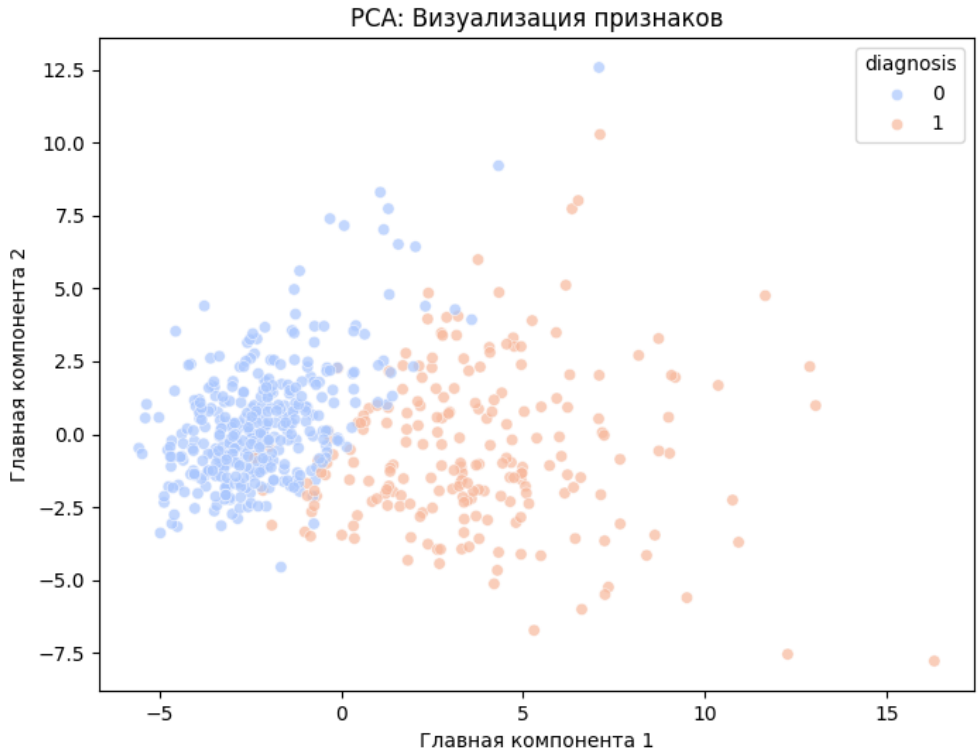

Рис.: Визуализация данных после PCA.

Логистическая регрессия

- Данные разделены на обучающую (70%) и тестовую (30%) выборки.
- Обучение модели на обучающих данных.
- Оценка модели на тестовой выборке:
  - **Точность:** 98.25%
  - Матрица ошибок:

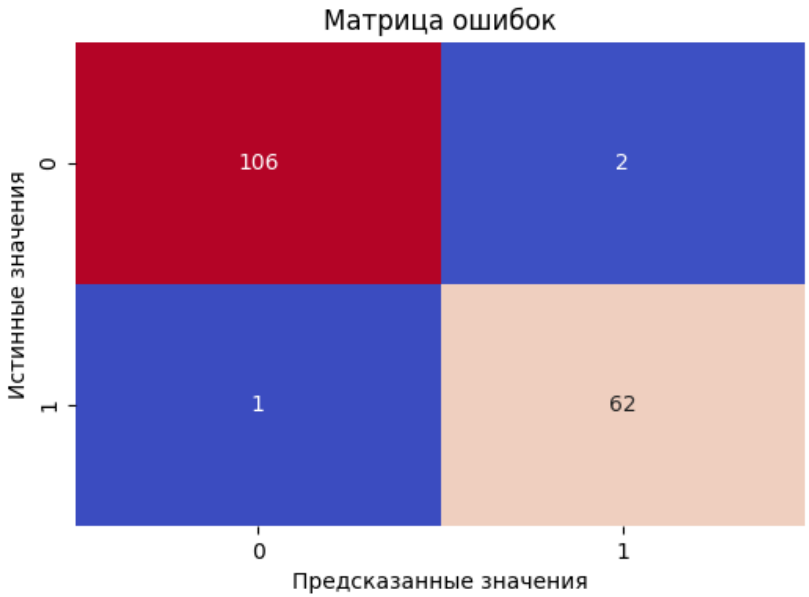

Рис.: Матрица ошибок.

Выводы

- Анализ главных компонент (PCA):
  - Уменьшение размерности данных позволило выделить ключевые компоненты, сохранив 98% дисперсии.
  - Визуализация данных в двумерном пространстве облегчила анализ структуры данных.
- Логистическая регрессия:
  - Модель показала высокую точность на тестовой выборке: **98.25%**.
  - Значения метрик:
    - Precision: **0.99** для класса 0, **0.97** для класса 1.
    - Recall: **0.98** для обоих классов.
    - F1-score: **0.99** для класса 0, **0.98** для класса 1.
  - Матрица ошибок показывает минимальное количество ошибок: 3 из 171 объектов.
- Итог:
  - Комбинация PCA и логистической регрессии доказала свою эффективность в задаче классификации опухолей молочной железы.
  - Подход может быть успешно применен к другим наборам данных с аналогичными характеристиками.