Практическое задание 4: Многомерный анализ данных для улучшения качества данных

Цель:

В этом задании вам предстоит использовать методы многомерного анализа данных для повышения качества набора данных путем изучения и обработки взаимодействий между переменными. Вы будете использовать тот же набор данных, что и в предыдущем задании, опираясь на улучшения, сделанные с помощью методов одномерного анализа. Задача состоит в выявлении многомерных зависимостей, отношений и причинно-следственных связей, что должно повысить прогностическую точность вашей модели.

Набор данных:

Продолжайте использовать набор данных, выбранный для предыдущего задания. Это обеспечит последовательность в оценке улучшений и отслеживание влияния различных методов на качество модели.

Инструкции:

1. **Первоначальный многомерный анализ:**
    - **Корреляционная матрица и диаграммы рассеяния:** выявите линейные и нелинейные зависимости между независимыми переменными, а также между независимыми и зависимыми переменными.
    - **Выявление мультиколлинеарности:** используйте анализ VIF (коэффициент инфляции дисперсии) для обнаружения мультиколлинеарности среди предикторов и сокращения избыточности путем выбора или преобразования переменных.
2. **Продвинутые методы многомерного анализа:**
    - **Анализ главных компонентов (PCA):** уменьшите размерность, если есть сильные корреляции между предикторами, сосредоточив внимание на захвате как можно большей дисперсии с меньшим количеством компонентов.
    - **t-SNE (t-распределение стохастического соседства):** визуализируйте данные высокой размерности в низкоразмерном пространстве, что полезно для обнаружения кластеров в сложных данных.
    - **UMAP (Универсальное приближение многообразия):** используйте UMAP для нелинейного уменьшения размерности, сохраняя локальную структуру данных.

- **Факторный анализ:** определите скрытые факторы, представляющие основные структуры в наборе данных, особенно если некоторые переменные концептуально связаны.
- **Кластерный анализ:** используйте методы кластеризации (например, k-means или иерархическую кластеризацию) для выявления естественных группировок в данных.

3. **Дополнительные методы создания признаков:**
   - **Генерация взаимодействий признаков:** создавайте взаимодействия между переменными для захвата нелинейных зависимостей.
   - **Полиномиальные признаки:** генерируйте полиномиальные комбинации признаков для моделирования нелинейных зависимостей.
   - **Биннинг и дискретизация:** преобразуйте непрерывные признаки в категориальные.
   - **Методы отбора признаков:** используйте рекурсивное исключение признаков (RFE) или регуляризационные методы (например, Lasso).
4. **Байесовские сети:**
   - **Обучение структуры:** используйте алгоритмы для обучения структуры Байесовской сети (например, Hill-Climbing).
   - **Условная независимость:** исследуйте условные зависимости для понимания влияния переменных друг на друга.
   - **Применение:** используйте сеть для отбора переменных, создания синтетических данных, балансировки классов.
5. **Повторное построение модели:**

После завершения многомерного анализа данных очистите, преобразуйте и сократите набор данных. Повторно обучите модель (регрессию или классификацию) из предыдущего задания, используя обновленные данные.

6. **Отчет:**

Ваш отчет должен включать следующие разделы:

- **Введение:** краткое описание набора данных, задачи и цели применения многомерного анализа.

- **Многомерный анализ данных:** описание процесса анализа и улучшения для каждого метода.
- **Результаты:** сравнение базовой модели, модели после одномерного и многомерного анализа.
- **Заключение:** краткое изложение выводов, влияние многомерного анализа на качество данных.

Сдача работы:

- Jupyter notebook с кодом анализа и построения модели.
- Отчет в PDF, подытоживающий анализ и выводы.