Вариант 5: Постройте модель, кластеризующую полученный датасет.

Датасет: «clustering\_1.csv»

1. Загрузить датасет в соответствии с вариантом и провести визуальный и разведочный анализ данных.
   1. Определить входные и выходные переменные.
   2. Проверить наличие пропущенных значений в данных.
   3. Для количественных признаков построить диаграммы рассеяния и диаграммы Box-and-Whisker, для категориальных признаков – гистограммы частот вариантов. Сделать выводы о характере распределений признаков, наличии выбросов.
   4. Для количественных переменных построить диаграммы рассеяния и рассчитать статистику (среднее значение, размах, число вариантов и др.).
2. Построить базовую модель машинного обучения.
   1. Устранить строки с пропущенными значениями.
   2. Провести OneHot-кодирование номинальных признаков (используя метод sklearn.preprocessing.OneHotEncoder с опцией drop=’first’).
   3. Провести OneHot-кодирование меток класса (для задач классификации).
   4. Визуализировать корреляционную матрицу признаков, сделать выводы.
   5. Разбить исходную выборку на обучающую и тестовую (70% первых строк включить в обучающую выборку, оставшиеся строки – в тестовую).
   6. Провести факторизацию обучающих данных с помощью метода главных компонент, получив 2 главных компоненты.
   7. Построить диаграмму рассеяния для точек в пространстве из двух главных компонент.
   8. Провести кластеризацию данных в пространстве двух главных компонентов на 3 кластера, используя метод k-средних и рассчитать коэффициент силуэта полученных кластеров.
3. Улучшить точность базовой модели.
   1. Выбрать метод и провести восстановление пропущенных значений в обучающей выборке.
   2. Предложить метод и устранить выбросные значения в обучающей выборке.
   3. Выбрать способ и провести преобразования входных и выходных переменных, привести обоснование выбранного способа преобразования.
   4. При необходимости провести отбор признаков / сократить размерность числа признаков/ сформировать новые признаки. Привести обоснование использованных методов формирования признаков.
   5. Выбрать модель и алгоритм машинного обучения для решения поставленной задачи или модифицировать базовую модель, привести обоснование выбора / модификаций.
   6. Провести обучение модели, рассчитать показатели качества модели на обучающей и тестовой выборках.
   7. Сравнить полученные результаты с аналогичными результатами для базовой модели, сделать выводы.
4. Оформить отчет по результатам проведенных исследований.
5. Пройти тестирование в своем личном кабинете