Задание III. «Кластеризация»

1. В файле «baseball.csv» находится выборка с информацией по игрокам в бейсбол, включая статистику их результативности, время участия в играх, лига, зарплата и т.д. Name (имя) нужно считать идентификатором записи. Загрузите этот файл и произведите следующие действия для кластерного анализа.

NAME	LABEL	ROLE	LEVEL
CrAtBat	Career Times at Bat	INPUT	INTERVAL
CrBB	Career Walks	INPUT	INTERVAL
CrHits	Career Hits	INPUT	INTERVAL
CrHome	Career Home Runs	INPUT	INTERVAL
CrRbi	Career BIs	INPUT	INTERVAL
CrRuns	Career Runs	INPUT	INTERVAL
Div	League and Division	INPUT	NOMINAL
Division	Division at the End of 1986	INPUT	NOMINAL
League	League at the End of 1986	INPUT	NOMINAL
logSalary	Log Salary	INPUT	INTERVAL
Name	Player's Name	ID	NOMINAL
nAssts	Assists in 1986	INPUT	INTERVAL
nAtBat	Times at Bat in 1986	INPUT	INTERVAL
nBB	Walks in 1986	INPUT	INTERVAL
nError	Errors in 1986	INPUT	INTERVAL
nHits	Hits in 1986	INPUT	INTERVAL
nHome	Home Runs in 1986	INPUT	INTERVAL
nOuts	Put Outs in 1986	INPUT	INTERVAL
nRBI	RBIs in 1986	INPUT	INTERVAL
nRuns	Runs in 1986	INPUT	INTERVAL
Position	Position(s) in 1986	INPUT	NOMINAL
Salary	1987 Salary in $ Thousands	INPUT	INTERVAL
Team	Team at the End of 1986	INPUT	NOMINAL
YrMajor	Years in the Major Leagues	INPUT	INTERVAL

2. Обработка пропусков. Переменная Salary (и log Salary) может содержать пропуски, произведите подстановку пропусков методом согласно вашему варианту. Пересчитайте logSalary как log(1+Salary), чтобы получить более симметричное распределение.
3. Нормализация переменных – приведите числовые переменные к близким шкалам с помощью методов для вашего варианта и закодируйте категориальные с помощью OneHotEncoder.
4. С помощью восходящей иерархической кластеризации с выбранными параметрами расстояния согласно вашему варианту постройте кластерную модель данных и дендрограмму для топ 20 кластеров.
5. Рассчитайте значение критерия pseudoF для вариантов кластеризации 2-20 кластеров, постройте график зависимости критерия от числа кластеров и выберите оптимальное (первый локальный пик критерия при обходе от малого числа кластеров к большому). Отметьте точку на графике. Сколько кластеров получилось?
6. С помощью метода проекции для вашего варианта постройте отображение на плоскость, цветом точки укажите номер кластера.
7. Выполните кластеризацию сферическими кластерами с прототипом методом из вашего варианта, также постройте проекцию как на шаге 6, определите наиболее типичного представителя (по имени) в каждом из кластеров.
8. Реализуйте шаги 3-7 в виде функции или класса.

9. Произведите дополнительную предобработку набора данных, сделав распределения переменных более симметричными. Для этого с помощью гисторамм или метода describe в dataframe или метода skew найдите переменные с одной модой и тяжелым правым хвостом, примените к ним преобразование log(1+x). Запустите функцию из шага 8. Как изменилось число кластеров, проекции и лучшие представители. Как считаете, субъективное качество кластеризации изменилось? Как и почему?
10. Отберите число наиболее значимых переменных из вашего варианта с помощью метода VarClus. Запустите функцию из шага 8. Как изменилось число кластеров, проекции и лучшие представители. Как считаете, субъективное качество кластеризации изменилось? Как и почему?

Задание IV. «Поиск аномалий»

11. «Творческое задание» на поиск аномалий. Загрузите файл mnist_small.csv. Данный набор данных содержит подмножество эталонного набора данных рукописных цифр MNIST. 5923 картинок 28x28 пикселей с изображением нуля и 76 картинок с изображением шестерки. Задача состоит в том, чтобы с использованием методов обучения без учителя для своего варианта построить одноклассовую модель на основе поиска аномалий, которая максимально хорошо отфильтрует шестерки (как аномалии) от нулей (как основной выборки). Признаки картинок описываются их координатами (в названии переменных, например «10x12») и значением яркости точки по этим координатам. Подбирая параметры метода и преобразуя признаки как посчитаете нужным, но не используя при этом информацию о label, постройте модель выявления аномалий с ERR меньше 0.2.
12. Постройте ROC кривую с ERR. Выведите 4 картинки с числами (28 на 28 пикселей):
    - самый типичный “0” – true negative с минимальной аномальностью
    - самая аномальная “6” – true positive с максимальной аномальностью
    - самый нетипичный “0” – false positive с максимальной аномальностью
    - самая неаномальная “6” – false negative с минимальной аномальностью

Запишите и перешлите для проверки JN реализующий шаги 1-10.

ВАРИАНТ	ПУНКТ 2	ПУНКТ 3	ПУНКТ 4	ПУНКТ 6	ПУНКТ 7	ПУНКТ 10	ПУНКТ 11
1	SimpleImputer (median)	StandardScaler	link=ward, dist=euclidean	AE	KMeans	7	SVM
2	SimpleImputer (mean)	RobustScaler	link=complete, dist=euclidean	tSNE	KMedoids	5	AE
3	KnnImputer (neighbors=3)	MinMaxScaler	link=complete, dist=manhattan	SOM	EM	7	KPCA
4	KnnImputer (neighbors=5)	MaxAbsScaler	link=average, dist=euclidean	PCA	KMeans	5	LOF
5	KnnImputer (neighbors=7)	Normalizer	link=average, dist=manhattan	NMF	KMedoids	7	KNN
6	SimpleImputer (median)	StandardScaler	link=ward, dist=euclidean	AE	EM	5	KNN
7	SimpleImputer (mean)	RobustScaler	link=complete, dist=euclidean	tSNE	KMeans	7	LOF
8	KnnImputer (neighbors=3)	MinMaxScaler	link=complete, dist=manhattan	SOM	KMedoids	5	SVM
9	KnnImputer (neighbors=5)	MaxAbsScaler	link=average, dist=euclidean	PCA	EM	7	KPCA
0	KnnImputer (neighbors=7)	Normalizer	link=average, dist=manhattan	NMF	KMeans	5	AE