**Лабораторная работа 1: Применение алгоритмов кластеризации для диагностики кризиса теплообмена в ЯЭУ.**

**Установка необходимых библиотек и инструментов**

Jupyter Notebook - это веб-приложение с открытым исходным кодом, которое вы можете использовать для создания и обмена документами, которые содержат живой код, уравнения, визуализацию, текст и научные исследования.

Jupyter Notebooks - это побочный проект проекта IPython, который раньше имел сам проект IPython Notebook. Название Jupyter происходит от основных поддерживаемых языков программирования, которые он поддерживает: Julia, Python и R. Jupyter поставляется с ядром IPython, которое позволяет писать свои программы на Python, но в настоящее время существует более 100 других ядер, которые вы можете использовать. также можно использовать.

pip install jupyter – установка Jupyter Notebook.

jupyter notebook – запуск Jupyter Notebook.

Numpy - это библиотека Python для вычислительно эффективных операций с многомерными массивами, предназначенная в основном для научных вычислений.

*import numpy as np*

Pandas - это библиотека Python, предоставляющая широкие возможности для анализа данных. С ее помощью очень удобно загружать, обрабатывать и анализировать табличные данные с помощью SQL-подобных запросов.

*import pandas as pd*

Основными структурами данных в Pandas являются классы Series и DataFrame. Первый из них представляет собой одномерный индексированный массив данных некоторого фиксированного типа. Второй - это двумерная структура данных, представляющая собой таблицу, каждый столбец которой содержит данные одного типа. Можно представлять её как словарь объектов типа Series.

**Выполнение лабораторной работы**

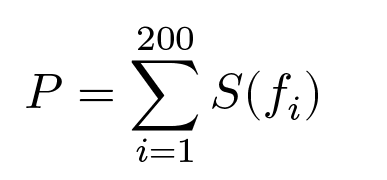
1. Прочтите данные из файлов varХ.csv, targetX.csv (где Х – номер варианта).

*Функции, которые могут пригодиться при решении: pd.read\_csv()*

1. Транспонируйте исходную матрицу. Каждый столбец будет спектром с 200 частотами.
2. Отобразите несколько первых и несколько последних записей.

*Функции, которые могут пригодиться при решении: .head(), .tail().*

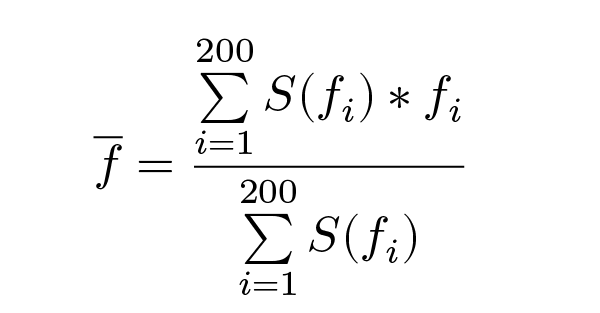
1. Постройте графики временных реализаций каждого спектра с помощью цикла или встроенных средств библиотеки pandas для визуализации. Используйте функцию plot() из библиотеки matplotlib.
2. Постройте временные реализации каждого спектра на одном графике. Используйте функцию plot() из библиотеки matplotlib..
3. Найдите мощность всего спектра по формуле



1. Выведите распределение целевой переменной targetX.csv. Зафиксируйте индекс начала кризиса теплообмена (значения равные 2).
2. Постройте график мощности спектра и начертите на нем вертикальную линию, которая будет разъединять участок без кризиса и с кризисом.
3. Выведите описательные статистики данных мощности спектра до кризиса (без кризиса) и с кризисом теплообмена.

*Функции, которые могут пригодиться при решении: .describe()*

1. Отдельно выведите в рабочую область средние значения мощности спектра до кризиса и после.
2. Постройте диаграммы ящиков с усами (boxplots) мощности спектра до кризиса и после на одном графики и сравните их. Опишите все наблюдения по построенным диаграммам.
3. Найдите среднюю частоту спектра, используя следующую формулу:



1. Постройте график значений средних частот спектра и начертите на нем вертикальную линию, которая будет разъединять участок без кризиса и с кризисом.
2. Выведите описательные статистики данных средних значений спектра до кризиса (без кризиса) и с кризисом теплообмена.

*Функции, которые могут пригодиться при решении: .describe()*

1. Отдельно выведите в рабочую область средние значения средних частот спектра до кризиса и после.
2. Постройте диаграммы ящиков с усами (boxplots) средних частот спектра до кризиса и после на одном графики и сравните их. Опишите все наблюдения по построенным диаграммам. Используется функцию boxplots() из библиотеки matplotlib.
3. Постройте график, где по оси Х будет отложена мощность, а по У отложена средняя частота. Раскрасьте точки на графике с помощью значений вектора целевой переменной.
4. Постройте 5 графиков с 5 парами (т.е. на каждом графике по 2) случайных частот, выбранных из исходного набора данных. Раскрасьте точки на графике с помощью значений вектора целевой переменной. Проведите прямую, равноудаленную от точек каждого класса (можно использовать МНК).
5. Примените метод понижения размерности (метод главных компонент) к исходному набору данных с частотами спектров. Визуализируйте 2 первые главные компоненты на плоскости и раскрасьте точки на графике с помощью значений вектора целевой переменной. (можно воспользоваться следующей библиотекой *from sklearn.decomposition import PCA)*. Не забудьте выполнить масштабирование многомерных данных перед их визуализацией на плоскости и понижением размерности.
6. Примените не менее 3-ех методов кластеризации (*KMeans, SpectralClustering, AgglomerativeClustering, DBSCA*N и др.) к исходным данным спектров и выполните их кластеризацию на 2 класса (без кризиса и с кризисом). Проверьте качество кластеризации по метрике homogeneity\_completeness\_v\_measure из библиотеки sklearn, модуля metrics (*from sklearn.metrics import homogeneity\_completeness\_v\_measure)*. Алгоритмы кластеризации можно найти в модуле *sklearn.clustering*.
7. Повысьте точность работы алгоритмов с помощью выбора информативных признаков (частот или спектров) из исходного набора данных. Заново примените алгоритмы и добейтесь наилучшей точности работы алгоритмов.
8. Оформите отчет по лабораторной работе в формате .ipynb с заголовками, комментариями, рисунками (с заголовками и названиями осей), ответами на контрольные вопросы, а также выводами о проделанной работе. Перед первым заголовком должно быть ваше ФИО и название группы. Назовите файл ФИО\_lab1.ipynb и сделайте файл .pdf с таким же названием, а затем сдайте оба файла преподавателю.

**Контрольные вопросы**

1. Какие существуют алгоритмы кластерного анализа данных? Назовите не менее 3-ех и опишите их суть с математической точки зрения и расскажите чем они отличаются друг от друга.
2. Какие метрики используются для оценки качества работы алгоритма кластеризации данных? Опишите данные метрики с математической точки зрения и скажите чем они отличаются друг от друга.
3. Каким способом можно повысить качество работы алгоритмов кластеризации?

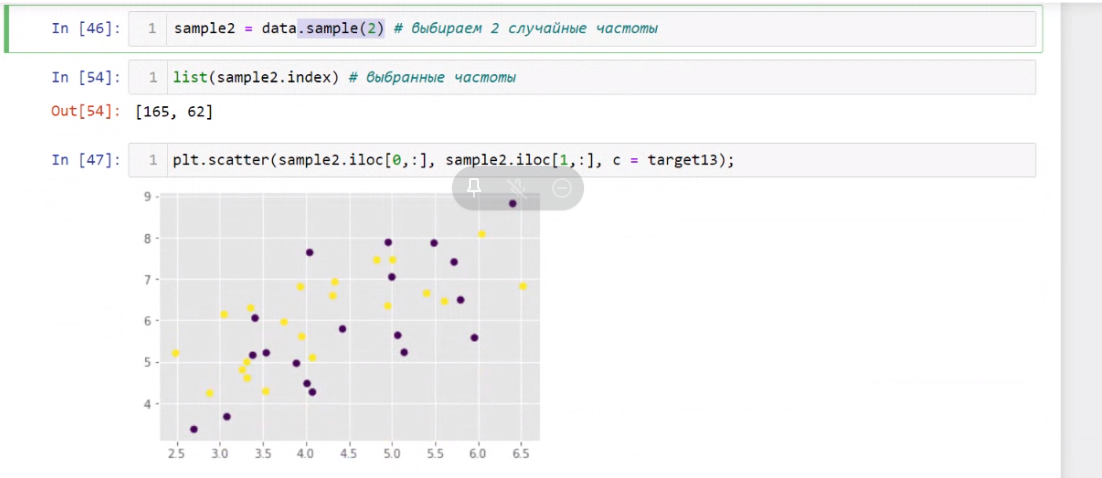
Погнали пояснялки

Чилситель range от 1 до 200

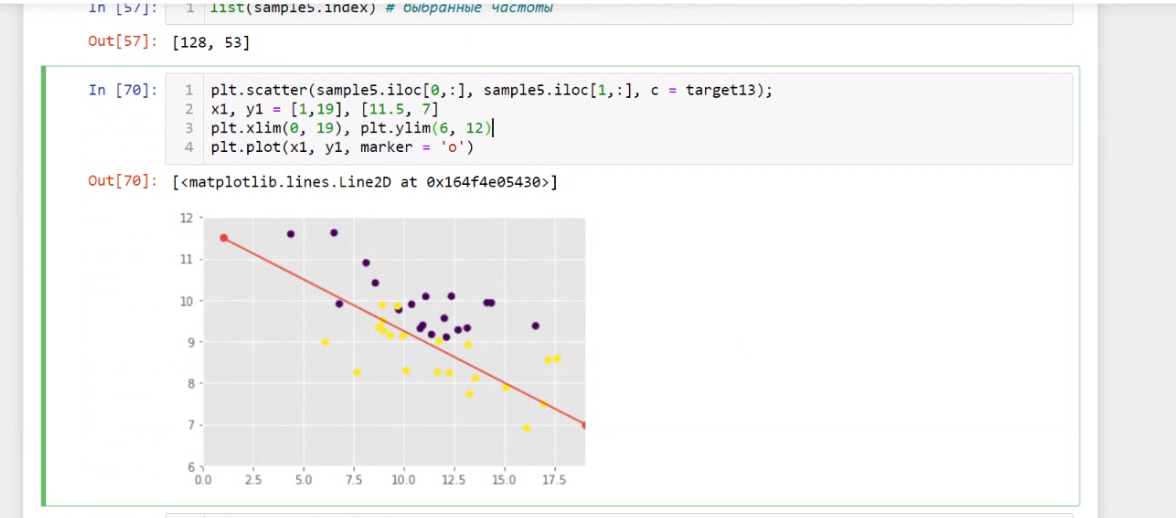
Np.arrange

Count plot.

В бокс

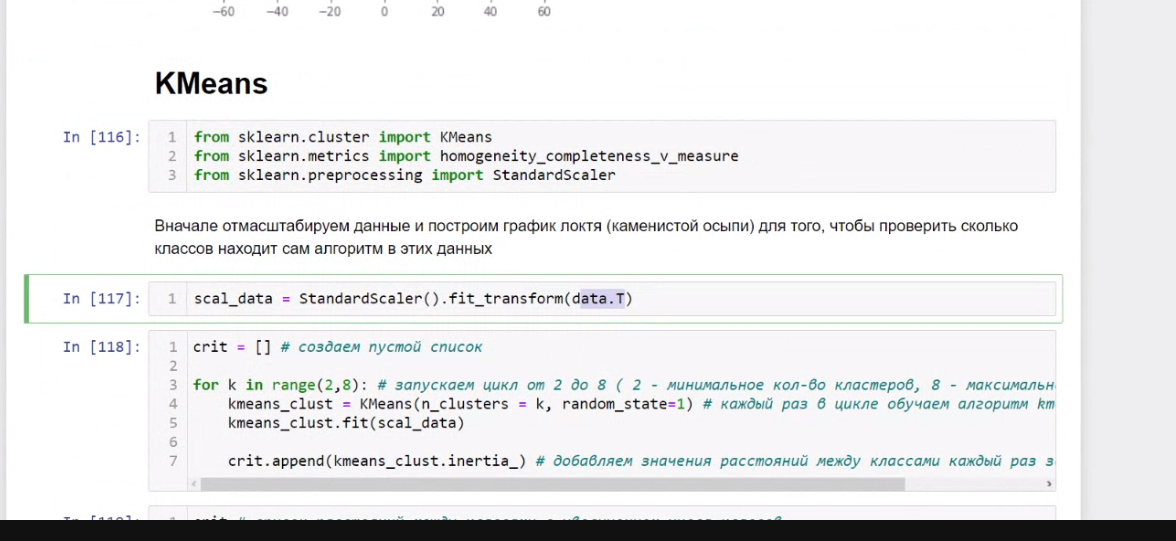


Для РСА добавить трансформ + визуализация РСА





Как выбрать информативные частоты. Надо найти такие признаки, Определить какие частоты влияют на кризис, а какие нет.



По вопросам – лучше с примером реаилаций

На лабу 2; тес сайз – 0.3

Best Estimator – его брать когда то, cv=5? Best\_estimator