# **Python и анализ данных**

Python является одним из самых популярных и востребованных языков программирования. В экосистеме этого языка существует огромное количество инструментов для решения разнообразных и специфичных задач. Big data – это структурированные или неструктурированные массивы данных большого объема. В python имеется множество пакетов, ориентированных на сферу Big data. В них имеются реализации сложных алгоритмов, которые упрощают написание кода.

Использование Python для анализа данных:

1. поиск данных,
2. обработка и моделирование данных,
3. визуализация данных.

Преимущества Python для анализа данных:

1. Отличное сообщество. Существует большое количество репозиториев с Python – проектами, благодаря чему почти всегда можно найти ответ на необходимый вопрос.
2. Легкость в изучении. Python – это один из самых простых языков для изучения благодаря своему простому синтаксису и читаемости.
3. Гибкость и масштабируемость. Гибкий и широкий набор инструментов позволяет использовать Python в самых разных отраслях.
4. Разнообразие библиотек. Возможность использовать огромное количество библиотек на разных этапах анализа данных.

Недостатки Python для анализа данных:

1. Разработка с использование динамической типизации проще и легче, но снижает скорость поиска ошибок в данных, связанных с разными типами.

# **NumPy (Numerical Python)**

NumPy – широко используемая библиотека пакетов Python, которая способна поддерживать специализированные структуры данных, к примеру, одномерные и двумерные массивы, которые расширяют возможности совершения различных вычислений.

Особенности NumPy:

* используется для выполнения как простых, так и значительно сложных расчетов.
* поддержка многомерных массивов.
* большое количество встроенных методов, которые применимы для вычислений на многомерных массивах.
* возможность совершать различные преобразования данных.
* поддержка не только числовых, но и других типов данных.

Сферы использования NumPy:

* создание новых массивных библиотек.

На основе этой библиотеки появляются новые типы массивов.

* Data Science.

В основе анализа данных лежит NumPy, которая используется на этапах работы с данными: извлечение и преобразование, анализ, моделирование и оценка, репрезентация.

* Machine Learning.

Некоторые библиотеки машинного обучения работают благодаря вычислительным мощностям NumPy.

* Визуализация данных.

NumPy позволяет визуализировать большие по размеру наборы данных.

# **Pandas**

Pandas – пакет - Python, который является вторым по популярности и использованию после NumPy. Библиотека Pandas построена на основе Numpy, благодаря чему у нее лучше производительность.

Особенности Pandas:

* поддержка объекта Dataframe, который предназначен для работы с индексированным массивом.
* этот пакет один из лучших для исследования данных.
* возможность использовать для работы с большими наборами данных: слияние и объединение данным, создание срезов и группировка данных.
* возможность работать с различными источниками данных, к примеру, с CSV и TSV – файлами.

Сферы использования Pandas:

* внесение изменений в существующий файл.

Импорт исходных электронных таблиц, совершение вычислений с ними, используя код, сохранение электронной таблицы с новыми результатами.

* помощь в визуализации данных.
* создание проектов машинного обучения.

Возможно объединить Pandas с другими библиотеками для выполнения задач машинного обучения.

# **Scikit – learn**

Самой функциональной и практичной библиотекой Python для ML является библиотека Scikit – learn, которая реализует все необходимые методы для задач классификации, кластеризации и регрессии.

Библиотеки NumPy, SciPy, matplotlib являются основоположниками Scikit – learn.

Эта библиотека удобна для использования, многофункциональна, популярна и к тому же находится в бесплатном доступе. Главным образом, в Scikit – learn, можно встретить огромное количество реализованных моделей машинного обучения такие, как деревья решений, ансамблевые модели, нейросети, линейные и метрические модели, модель наивного Байеса и прочие. Вместе с алгоритмами кластеризации, классификации и регрессии, в библиотеке разработано множество методов для уменьшения размерности, и обработки данных, которые включают стандартизацию и нормализацию, OrdinalEncoder и OneHotEncoder, векторизацию, бинаризацию и преобразование непрерывных данных.

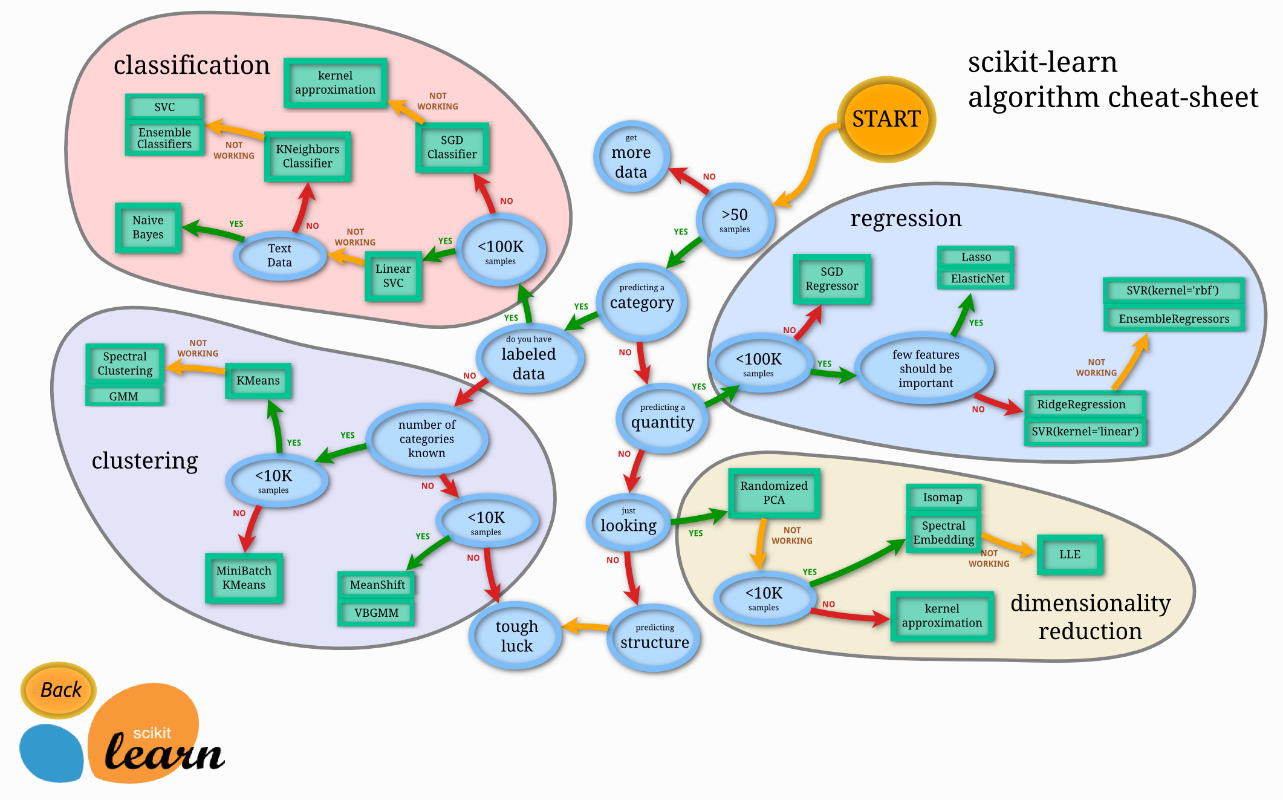


Рисунок 3 - Scikit - learn

Несмотря на широкий спектр преимуществ, Scikit – learn не может обеспечить поддержку обучения ассоциативным правилам и с подкреплением, сетей Кохонена.