**xStart1**

Исследовательский анализ данных (Exploratory data analysis, EDA) используется для анализа и исследования наборов данных и обобщения их основных характеристик

EDA помогает определить, как лучше всего манипулировать источниками данных для получения необходимых ответов, упрощая специалистам по обработке и анализу данных обнаружение закономерностей, выявление аномалий, проверку гипотез или предположений.

EDA в основном используется для того, чтобы увидеть, что данные могут выявить за пределами формального моделирования или проверки гипотез, и обеспечивает лучшее понимание переменных набора данных и отношений между ними. Это также может помочь определить, подходят ли статистические методы, которые вы рассматриваете для анализа данных.

Основная цель EDA — помочь взглянуть на данные, прежде чем делать какие-либо предположения. Это может помочь выявить очевидные ошибки, а также лучше понять закономерности в данных, обнаружить выбросы или аномальные события, найти интересные связи между переменными.

Специалисты по обработке и анализу данных могут использовать исследовательский анализ, чтобы убедиться, что результаты, которые они получают, являются достоверными и применимыми к любым желаемым бизнес-результатам и целям. EDA также помогает заинтересованным сторонам, подтверждая, что они задают правильные вопросы. EDA может помочь ответить на вопросы о стандартных отклонениях, категориальных переменных и доверительных интервалах.

Конкретные статистические функции и методы, которые можно выполнять с помощью инструментов EDA, включают:

* Методы кластеризации и уменьшения размерности, которые помогают создавать графические представления многомерных данных, содержащих множество переменных.
* Одномерная визуализация каждого поля в необработанном наборе данных со сводной статистикой.
* Двумерные визуализации и сводная статистика, которые позволяют оценить связь между каждой переменной в наборе данных и целевой переменной, которую вы просматриваете.
* Многомерные визуализации для сопоставления и понимания взаимодействий между различными полями данных.
* Кластеризация K-средних — это метод кластеризации в [обучении без учителя](https://developer.ibm.com/articles/cc-unsupervised-learning-data-classification), при котором точки данных распределяются по K-группам, т.е. по количеству кластеров, в зависимости от расстояния от центроида каждой группы. Точки данных, ближайшие к определенному центроиду, будут объединены в одну категорию. Кластеризация K-средних обычно используется для сегментации рынка, распознавания образов и сжатия изображений.
* Прогностические модели, такие как линейная регрессия, используют статистику и данные для прогнозирования результатов.

Существует четыре основных типа EDA:

* **Одномерный неграфический.** Это простейшая форма анализа данных, при которой анализируемые данные состоят всего из одной переменной. Поскольку это одна переменная, она не имеет отношения к причинам или отношениям. Основной целью одномерного анализа является описание данных и поиск закономерностей, которые в них существуют.
* **Одномерная графика.** Неграфические методы не дают полной картины данных. Поэтому требуются графические методы. К распространенным типам одномерной графики относятся:
  + Штамбовые и листовые графики, которые показывают все значения данных и форму распределения.
  + Гистограммы, гистограмма, на которой каждый столбец представляет частоту (количество) или долю (количество/общее количество) вариантов для диапазона значений.
  + Ящичковые диаграммы, которые графически изображают сводку из пяти чисел минимума, первого квартиля, медианы, третьего квартиля и максимума.
* **Многомерные неграфические:**Многомерные данные возникают из нескольких переменных. Многомерные неграфические методы САПР обычно показывают взаимосвязь между двумя или более переменными данных с помощью перекрестной таблицы или статистики.
* **Многовариантная графика:** Многомерные данные используют графику для отображения связей между двумя или более наборами данных. Наиболее часто используемым рисунком является сгруппированная линейчатая диаграмма или линейчатая диаграмма, где каждая группа представляет один уровень одной из переменных, а каждая полоса в группе представляет уровни другой переменной.

К другим распространенным типам многомерной графики относятся:

* Точечная диаграмма, которая используется для отображения точек данных по горизонтальной и вертикальной оси, чтобы показать, насколько одна переменная подвержена влиянию другой.
* Многомерная диаграмма, представляющая собой графическое представление взаимосвязей между факторами и реакцией.
* Запустите диаграмму, которая представляет собой линейный график данных, построенный во времени.
* Пузырьковая диаграмма, представляющая собой визуализацию данных, отображающую несколько кругов (пузырьков) на двумерном графике.
* Тепловая карта, представляющая собой графическое представление данных, где значения изображены цветом.

[101] (Link: <https://www.ibm.com/topics/exploratory-data-analysis>)

**xStop1**

**xStart2**

### Краткое описание переменных анализа

📜 Набор данных для анализа

* **Age**: age of the patient [years]
* **Sex**: sex of the patient [M: Male, F: Female]
* **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
* **RestingBP**: resting blood pressure [mm Hg]
* **Cholesterol**: serum cholesterol [mm/dl]
* **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
* **RestingECG**: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
* **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
* **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
* **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
* **ST\_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
* **HeartDisease**: output class [1: heart disease, 0: Normal]

Наш набор данных содержит в общей сложности 6 числовых переменных:

***Age, RestingBP, Cholesterol, MaxHR, Oldpeak, HeartDisease***

В нашем наборе данных имеется также 6 категориальных переменных:

***Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST\_Slope***

**xStop2**

**xStart3**

Категориальные значения — это тип данных, которые могут быть сгруппированы в различные категории, такие как пол, цвет или жанр. Числовые значения — это данные, которые имеют числовое значение, например возраст, рост или цена. Нам необходимо преобразовать категориальные значения в числовые в САПР по нескольким причинам:

* Требования к моделированию: Если вы планируете создавать прогнозные модели или модели машинного обучения, необходимы числовые представления. Многие алгоритмы машинного обучения, особенно основанные на математических уравнениях, работают с числовыми входными данными.
* Визуализация: Числовые данные часто легче визуализировать и интерпретировать. Графики, диаграммы и другие визуализации обычно используются в САПР для получения аналитических сведений о данных, а числовые представления облегчают создание осмысленных визуализаций данных.
* Статистический анализ: Некоторые статистические тесты и анализы предполагают наличие числовых данных. Например, коэффициенты корреляции, регрессионный анализ и другие статистические методы предназначены для работы с числовыми переменными.
* Согласованность типов данных: преобразование категориальных переменных в числовые представления помогает поддерживать согласованность типов данных в наборе данных. Такая согласованность может упростить процесс анализа данных и сделать его более понятным.
* Совместимость с методами анализа: Многие статистические методы и методы машинного обучения требуют числовых входных данных. Преобразование категориальных переменных в числовые представления позволяет применять к набору данных более широкий спектр методов анализа.

Существует два распространенных метода преобразования категориальных переменных в числовые:

1. One-Hot Encoding: Этот метод создает двоичные столбцы для каждой категории, представляющие наличие или отсутствие этой категории.
2. Кодировка метки: Этот метод присваивает каждой категории уникальное целое число.

Однако горячее кодирование более распространено для двоичных категориальных переменных, так как оно явно представляет каждую категорию независимо.

Здесь мы используем метод One-Hot Encoding для преобразования данных.

**xStop3**

**xStart4**

**bbb**

**xStop4**

**xStart5**

**ccc**

**xStop5**

**xStart6**

**ddd**

**xStop6**

**xStart7**

**eee**

**xStop7**

**xStart8**

**fff**

**xStop8**

**xStart9**

**hhh**

**xStop9**

**xStart10**

**iii**

**xStop10**

**xStart11**

EDA на наборе данных показала нам, как каждая переменная связана с переменной отклика и как мы можем сделать нашу модель эффективной, используя различные методы EDA. Визуализации данных поднимают наше понимание набора данных на более высокий уровень, позволяя нам делать выводы.

Интеграция классификатора дерева принятия решений в наш анализ расширяет прогностические возможности нашей модели. В этом блоге мы не только изучили набор данных с помощью методов EDA, но и сделали еще один шаг вперед, внедрив модель машинного обучения. Такой целостный подход позволяет нам использовать сильные стороны как статистического анализа, так и прогнозного моделирования, способствуя более глубокому пониманию сложной динамики, связанной с экстремальными погодными явлениями.

**xStop11**

**xStart12**

**aaa**

**xStop12**

**xStart13**

**aaa**

**xStop13**

**xStart14**

**bbb**

**xStop14**

**xStart15**

**ccc**

**xStop15**

**xStart16**

**ddd**

**xStop16**

**xStart17**

**eee**

**xStop17**

**xStart18**

**fff**

**xStop18**

**xStart19**

**hhh**

**xStop19**

**xStart20**

**iii**

**xStop20**

**xStart21**

**iii**

**xStop21**

**xStart22**

**aaa**

**xStop22**

**xStart23**

**aaa**

**xStop23**

**xStart24**

**bbb**

**xStop24**

**xStart25**

**ccc**

**xStop25**

**xStart26**

**ddd**

**xStop26**

**xStart27**

**eee**

**xStop27**

**xStart28**

**fff**

**xStop28**

**xStart29**

**hhh**

**xStop29**

**xStart30**

**iii**

**xStop30**

**xStart31**

**iii**

**xStop31**

**xStart32**

**aaa**

**xStop32**

**xStart33**

**aaa**

**xStop33**

**xStart34**

**bbb**

**xStop34**

**xStart35**

**ccc**

**xStop35**

**xStart36**

**ddd**

**xStop36**

**xStart37**

**eee**

**xStop37**

**xStart38**

**fff**

**xStop38**

**xStart39**

**hhh**

**xStop39**

**xStart40**

**iii**

**xStop40**

**xStart41**

**iii**

**xStop41**

**xStart42**

**aaa**

**xStop42**

**xStart43**

**aaa**

**xStop43**

**xStart44**

**bbb**

**xStop44**

**xStart45**

**ccc**

**xStop45**

**xStart46**

**ddd**

**xStop46**

**xStart47**

**eee**

**xStop47**

**xStart48**

**fff**

**xStop48**

**xStart49**

**hhh**

**xStop49**

**xStart50**

**iii**

**xStop50**

**xStart51**

**xStop51**

**xStart52**

EDA на

**xStop52**

**xStart53**

EDA на наборе

**xStop53**

**xStart54**

**xStop54**

**xStart55**

EDA на наборе данных показала нам, как каждая переменная связана с переменной отклика и как мы можем сделать нашу модель эффективной, используя различные методы EDA. Визуализации данных поднимают наше понимание набора данных на более высокий уровень, позволяя нам делать выводы.

Интеграция классификатора дерева принятия решений в наш анализ расширяет прогностические возможности нашей модели. В этом блоге мы не только изучили набор данных с помощью методов EDA, но и сделали еще один шаг вперед, внедрив модель машинного обучения. Такой целостный подход позволяет нам использовать сильные стороны как статистического анализа, так и прогнозного моделирования, способствуя более глубокому пониманию сложной динамики, связанной с экстремальными погодными явлениями.

**xStop55**

**xStart56**

EDA на наборе данных показала нам, как каждая переменная связана с переменной отклика и как мы можем сделать нашу модель эффективной, используя различные методы EDA. Визуализации данных поднимают наше понимание набора данных на более высокий уровень, позволяя нам делать выводы.

Интеграция классификатора дерева принятия решений в наш анализ расширяет прогностические возможности нашей модели. В этом блоге мы не только изучили набор данных с помощью методов EDA, но и сделали еще один шаг вперед, внедрив модель машинного обучения. Такой целостный подход позволяет нам использовать сильные стороны как статистического анализа, так и прогнозного моделирования, способствуя более глубокому пониманию сложной динамики, связанной с экстремальными погодными явлениями.

**xStop56**

**xStart57**

EDA на наборе данных показала нам, как каждая переменная связана с переменной отклика и как мы можем сделать нашу модель эффективной, используя различные методы EDA. Визуализации данных поднимают наше понимание набора данных на более высокий уровень, позволяя нам делать выводы.

Интеграция классификатора дерева принятия решений в наш анализ расширяет прогностические возможности нашей модели. В этом блоге мы не только изучили набор данных с помощью методов EDA, но и сделали еще один шаг вперед, внедрив модель машинного обучения. Такой целостный подход позволяет нам использовать сильные стороны как статистического анализа, так и прогнозного моделирования, способствуя более глубокому пониманию сложной динамики, связанной с экстремальными погодными явлениями.

**xStop57**