Исследовательский анализ включает в себя изучение данных и поиск связей между переменными, которые ранее были неизвестны. Вот что вам нужно знать:

• EDA помогает обнаруживать взаимосвязи между показателями в данных, которые не являются доказательством существования корреляции, как обозначается фразой «Корреляция не подразумевает причинно-следственную связь».

• Это полезно для обнаружения новых связей и формирования гипотез. Он управляет планированием проектирования и сбором данных.

Пример исследовательского анализа

Изменение климата становится все более важной темой, поскольку глобальная температура постепенно повышается на протяжении многих лет. Один из примеров исследовательского анализа данных об изменении климата включает в себя повышение температуры за период с 1950 по 2020 год, а также рост человеческой деятельности и индустриализацию, чтобы найти взаимосвязи на основе данных. Например, вы можете увеличить количество заводов, автомобилей на дорогах и самолетов, чтобы увидеть, как это коррелирует с повышением температуры.

Исследовательский анализ исследует данные для поиска взаимосвязей между показателями без выявления причины. Это наиболее полезно при формулировании гипотез.

Для предварительной обработки набора данных KAMIR-NIH мы классифицировали все атрибутивные признаки по различным категориям, например, категориальные признаки, непрерывные признаки и дискретные признаки. Мы определили различные правила предварительной обработки для этих различных типов атрибутов. Для категориальных переменных мы применили кодирование меток [32], а также одно горячее кодирование [26] для предварительной обработки этих переменных. Для непрерывных атрибутов мы классифицировали набор данных по диапазонам, а затем применили кодировку меток для этих определенных подклассов. Для некоторых категориальных и непрерывных переменных, содержащих несколько значений, мы применили одно горячее кодирование, чтобы упростить управление значениями этих атрибутов. Одно горячее кодирование является одним из лучших решений для управления несколькими значениями и предварительной обработки тех атрибутов, которые содержат более одного параметра. В нашем наборе данных также есть атрибуты с двоичным значением. Для таких атрибутов, содержащих ровно два значения, мы преобразовали их в двоичную форму (0 и 1), обозначив 0 как No, 1 как Yes.

В нашем датасете было много атрибутов, которые не нужны для применения различных алгоритмов. Чтобы сделать наши данные более конкретными и безошибочными, мы удалили эти атрибуты из нашего набора данных. Например, для некоторых атрибутов типа даты, содержащих дату и время, нет необходимости использовать эти атрибуты в обучающих моделях. Поэтому мы удалили эти атрибуты. В случае с нашим набором данных некоторые атрибуты отсутствовали в нашем наборе данных, и они очень важны для моделей прогнозирования. На основе нашего текущего набора данных мы вывели эти атрибуты, используя другие атрибуты, и классифицировали их для использования этих атрибутов. Например, артериальное давление (АД), индекс массы тела (ИМТ), общий холестерин и частота сердечных сокращений (ЧСС) отсутствовали в наборе данных, но они были необходимы для моделей прогнозирования. Мы вывели эти атрибуты из других атрибутов и классифицировали их соответствующим образом.

Мы также следовали рекомендациям Корейского общества гипертонии [33, 34] по категоризации артериального давления, а затем применяли кодирование меток для преобразования данных. ИМТ рассчитывается как выражение кг/м2 от веса (кг) и роста пациента (м), а затем применили корейские стандарты [35] для категоризации значений ИМТ.Мы применили Национальные рекомендации по лечению холестерина [36] для классификации липопротеинов низкой плотности (ЛПНП), липопротеинов высокой плотности (ЛПВП) и общего холестерина у корейских пациентов. Предпочтительный уровень триглицеридов — менее 150 мг/дл (1,7 ммоль/л), повышенный пограничный — 150–199 мг/дл (1,7–2,2 ммоль/л), повышенный уровень — 200–499 мг/дл (2,3–5,6 ммоль/л), а очень высокий уровень триглицеридов — 500 мг/дл (5,6 ммоль/л) или выше [37]. Вне отечественных стандартов, согласно критериям ВОЗ, окружность бедер и талии является индикатором для диагностики абдоминального ожирения [38] и указывает на абдоминальное ожирение при WHR>0,9 для мужчин и >0,85 для женщин.  
С−реактивный белок (высокочувствительный СРБ, hs−СРБ) использовался в качестве предиктора сердечно-сосудистого риска у здоровых взрослых [39]. Люди с высокими значениями hs−CRP имеют высокий риск развития острого коронарного синдрома, а люди с низкими значениями имеют низкий риск. У людей с более высокими результатами hs−CRP в верхнем диапазоне нормы риск сердечного приступа примерно в 1,5-4 раза выше, чем у людей с более низкой частотой. Корейское общество диагностической радиологии использует те же критерии, что и Американская кардиологическая ассоциация [39, 40] и Центры по контролю и профилактике заболеваний США. Эти значения являются частью общего процесса оценки острого коронарного синдрома. Высокий уровень сахара в крови означает в основном диабет. Тем не менее, многие заболевания и системные состояния, кроме диабета, могут повышать уровень сахара в крови. Ниже приведена краткая информация о значении каждого результата теста. Это резюме основано на данных Американской диабетической ассоциации и классифицируется по нормальному уровню глюкозы натощак, преддиабетической стадии и диабету [38]. Концентрация креатинина в сыворотке крови повышается при нарушении функции почек. Аномальный диапазон для мужчин составляет >1,2 мг/дл и >1,0 мг/дл для женщин [41].

В нашем наборе данных мы также разобрались с отсутствующими значениями. В медицинском наборе данных очень сложно работать с пропущенными значениями, особенно когда данные очень чувствительны. Неправильное и неадекватное обращение с пропущенными значениями приведет к низкому прогнозированию фактора риска и наоборот [42]. Когда мы применяли алгоритмы машинного обучения для прогнозирования риска, ранней диагностики и прогнозирования острого коронарного синдрома, мы использовали различные методы импутации для нормализации данных, например, импутирование среднего значения [43, 44] и вменение k-ближайших соседей (k-NN) [43].  
В ходе предварительной обработки данных мы выяснили, что некоторые пациенты перенесли множественные сердечные события. Таким образом, мы классифицировали пациентов, перенесших несколько сердечных событий, в одно сердечное событие в зависимости от тяжести, осложнений и эффективности этого события. Например, пациент уже сделал АКШ, а потом умер из-за сердечно-сосудистых заболеваний, мы внесли этого пациента в КК, а не в АКШ.