Мы используем датасет по острому коронарному синдрому под названием Korea Acute Myocardial Infarction Registry (KAMIR-NIH) [11], который зарегистрирован в 52 больницах Кореи и содержит данные обо всех пациентах с ноября 2011 года по декабрь 2019 года. В своей исследовательской работе мы используем двухлетний набор данных, содержащий 551 различных атрибутов и медицинские карты 13 104 пациентов с двухлетним наблюдением после выписки из стационара. Однако существует ограничение на передачу этих данных, поскольку данные являются конфиденциальными и недоступны публично. Подробная информация о реестре размещена на сайте КАМИР (http://www.kamir.or.kr). В нашей выборке данных у нас есть вся основная медицинская информация о пациентах, такая как возраст, артериальное давление, частота сердечных сокращений, рост, вес, другие заболевания, а также предыдущая медицинская карта пациентов, страдающих каким-либо другим заболеванием или уже имеющих сердечную недостаточность, или какова степень тяжести состояния пациента. У нас также есть полные записи о приеме лекарств для пациентов с сердечными заболеваниями с двухлетним наблюдением. В данной работе основные неблагоприятные сердечно-сосудистые события (MACE) определяются как сердечная смерть (БК), несердечная смерть (НИЗ), инфаркт миокарда (ИМ), повторное чрескожное коронарное вмешательство (ре-ЧКВ) и аортокоронарное шунтирование (АКШ).  
2.3 Извлечение данных

Извлечение данных — это процесс извлечения или извлечения данных из неструктурированных или полуструктурированных источников данных для дальнейшей обработки данных для достижения требуемых результатов. В случае с набором данных KAMIR-NIH он находится в необработанном виде и содержит противоречивые, зашумленные и неполные данные. Он также содержит избыточность данных и выбросы. Чтобы решить эти проблемы, мы предварительно обрабатываем эти данные в понятном формате, чтобы получить больше полезной информации. Мы должны применять методы извлечения данных для извлечения и манипулирования важными признаками и записями из всего набора данных. Прежде всего, мы удалили атрибуты даты из набора данных KAMIR-NIH, поскольку эти атрибуты не влияют на раннюю диагностику и прогноз основных неблагоприятных сердечно-сосудистых событий. Во-вторых, из датасета были исключены все атрибуты, содержащие информацию о препаратах для пациентов, поскольку эти атрибуты не являются обязательными для требуемых результатов и содержат более 70% нулевых значений. В-третьих, все атрибуты, содержащие более 70% значений NULL, удаляются из набора данных. Удалив всю эту ненужную информацию из датасета, мы извлекли важные данные из датасета. Извлечение данных проиллюстрировано в Рис 3 в котором мы использовали набор данных KAMIR-NIH (N = 13 104) и исключили всех пациентов, умерших в больнице во время госпитализации (Excluded N = 504). Мы также исключили пациентов, которые не проводили двухлетнее наблюдение (исключено N = 1,411). После исключения всех ненужных данных из набора данных KAMIR-NIH у нас были пациенты с острым коронарным синдромом, которые были живы в течение двухлетнего наблюдения после выписки из больницы (N = 11 189). Затем этот набор данных был разделен на наборы данных «Инфаркт миокарда с подъемом сегмента ST» (STEMI) (N = 5,389) и «Инфаркт миокарда без подъема сегмента ST» (NST) (N = 5,800), а затем разделен на обучающие данные (70%) и данные тестирования (30%). Полные процессы извлечения данных проиллюстрированы в Рис 3.