

Технологии организации, обработки и хранения статистических данных

ФИО преподавателя: Митина О.А.

e-mail: alogmi@yandex.ru

3

Лекция

Методология CRISP-DM

Условия обучения

- По итогам изучения дисциплины проводится экзамен
- В течение семестра необходимо выполнить все практические работы

Методология CRISP-DM



Модель процесса CRISP-DM

Методология CRISP-DM

1 Понимание бизнеса



☐ Определить бизнес-цели

☐ Оценить ситуацию

☐ Определить цели анализа данных

☐ Составить план проекта

2 Понимание данных



☐ Собрать исходные данные

☐ Описать данные

☐ Исследовать данные

☐ Проверить качество данных

3 Подготовка данных



☐ Отобрать данные

☐ Очистить данные

☐ Получить производные данные

☐ Объединить данные

☐ Перевести данные в нужный формат

Методология CRISP-DM

3 Моделирование

☐

- ☐ Выбрать алгоритм
- ☐ Построить модель
- ☐ Оценить модель
- ☐ Протестировать модель

4 Оценка

☐

- ☐ Оценить результаты
- ☐ Провести аудит всех шагов моделирования
- ☐ Определить следующие шаги

3 Развертывание

☐

- ☐ Запланировать развертывание
- ☐ Запланировать поддержку проекта
- ☐ Подготовить документацию
- ☐ Провести аудит проекта

Формы представления данных

неструктурированные;
структурированные;
слабоструктурированные.

Формы представления данных

Столбцы (поля, колонки, переменные, атрибуты, признаки)

Строки
(записи,
прецеденты,
примеры)

8

Формы представления данных

390045 г. Рязань, ул. Ленина, д. 45 корп. 1



| Поле | Значение |
|--------|----------|
| Индекс | 390045 |
| Город | Рязань |
| Улица | Ленина |
| Дом | 45 |
| Корпус | 1 |

Формы представления данных

| | |
|-------------|---------------------|
| числовой; | шкала наименований; |
| символьный; | шкала порядка; |
| логический; | интервальная шкала; |
| дата/время. | шкала отношений. |

Номинальные переменные

| Переменная | Категории |
|-------------------|--|
| Наличие машины | Да Нет |
| Кредитная история | Положительная Отрицательная Нет данных |
| Провайдер | Мегафон МТС Билайн |

Ординальные переменные

| Переменная | Категории |
|-------------------|--|
| Наличие машины | Да Нет |
| Кредитная история | Положительная Отрицательная Нет данных |
| Провайдер | Мегафон МТС Билайн |

Соответствие между типами и видами данных

| Тип данных | Вид данных | |
|------------|-------------|------------|
| | Непрерывный | Дискретный |
| Числовой | + | + |
| Строковый | | + |
| Логический | | + |
| Дата/время | + | + |

Пример упорядоченных наборов данных

| Дата | Количество | Сумма |
|------------|------------|--------|
| 01.02.2017 | 4 | 283,31 |
| 01.02.2017 | 1 | 72,48 |
| 01.02.2017 | 1 | 173,32 |
| 02.02.2017 | 6 | 294,84 |
| 02.02.2017 | 2 | 405,76 |
| 02.02.2017 | 12 | 303,13 |
| 02.02.2017 | 1 | 210,50 |
| 03.02.2017 | 6 | 512,16 |
| 03.02.2017 | 3 | 156,96 |

Пример неупорядоченных наборов данных

| Номер | Банк | Город | Филиалы | Собственные активы |
|-------|--------------|-----------------|---------|--------------------|
| 2 | Внешторгбанк | Москва | 32 | 23236327 |
| 3 | Газпромбанк | Москва | 27 | 9255041 |
| 4 | Альфа-Банк | Москва | 17 | 12446938 |
| 5 | ОАО «ПСБ» | Санкт-Петербург | 44 | 1275859 |
| 6 | Банк Москвы | Москва | 34 | 3335734 |
| 7 | АКБ «ДИБ» | Москва | 0 | 261 6993 |

Одна транзакция



| Код транзакции | Товар |
|----------------|--------------------|
| 10200 | Йогурт «Чудо» 0,4 |
| 10200 | Батон «Рязанский» |
| 10201 | Вода «Боржоми» 0,5 |
| 10201 | Сахарный песок |

Особенности бизнес-данных

Редко накапливаются специально
для задач анализа;
Содержат ошибки, выбросы,
противоречия, пропуски;
Объем данных велик.

Формализация данных - принципы

1. Абстрагироваться от существующих информационных систем и имеющихся в наличии данных.
2. Описать все факторы, потенциально влияющие на анализируемый процесс/объект.
3. Экспертно оценить значимость каждого фактора.
4. Определить способ представления информации – число, дата, да/нет, категория (т. е. тип данных).
5. Собрать легкодоступные факторы .
6. Оценить сложность и стоимость сбора средних и наименее важных по значимости факторов.

Информативность данных

| Признак |
|---------|
| 1 |
| 1 |
| 1 |
| 1 |
| (1) |

| Признак |
|---------|
| 1 |
| 1 |
| 0 |
| 1 |
| (2) |

| № паспорта |
|-------------|
| 0936-866096 |
| 8355-512928 |
| 8017-098418 |
| 0094-732300 |
| (3) |

| Пол | Gender |
|-----|--------|
| Жен | 0 |
| Жен | 0 |
| Муж | 1 |
| Муж | 1 |
| (4) | |

Примеры неинформативных данных

Сбор данных. Методы сбора данных

1. Получение из учетных систем.
2. Получение данных из косвенных источников информации.
3. Использование открытых источников.
4. Приобретение данных у специализированных компаний.
5. Проведение собственных мероприятий по сбору данных.
6. Ввод вручную.

Выводы

1. Методология CRISP-DM .
2. Формы представления данных.
3. Типы данных.
4. Особенности бизнес-данных.
5. Информативность данных.
6. Методы сбора данных. Подготовка данных.

Список литературы

- Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. – М.: МЦНМО, 2016. – 368 с.
- Мхитарян В.С. Анализ данных: учебник для академического бакалавриата / под ред. В.С. Мхитаряна. – М.: Изд. Юрайт, 2017 – 490 с.
- Хрусталёв Е.М. Агрегация данных в OLAP-кубах. [http:// www . olap . ru /](http://www.olap.ru/)

Темы дисциплины

- 1 Анализ данных. Основные понятия и определения
- 2 Бизнес-аналитика. Основные понятия и определения
- 3 Методология CRISP-DM
- 4 Многомерная модель данных
- 5-6 Интеграция данных и бизнес-аналитика
- 7-8 Интеграция данных
- 9 Хранилища данных
- 10 Процессы информативной корпоративной фабрики
- 11 Базовые архитектуры корпоративной информационной фабрики
- 12 Технология OLAP и ее особенности
- 13 Понятие OLAP-куба. Операции над OLAP-кубами
- 14 Аналитические платформы. Инструменты бизнес-аналитики
- 15-16 Большие данные. Наука о данных

Спасибо за внимание!