**Шаблоны для документации результатов дипломной работы**

CRISP-DM является проверенным в отрасли способом руководства действиями по интеллектуальному анализу данных. Как методология, она включает в себя описание типичных этапов проекта, задач и вовлекается в каждую фазу, и объяснение взаимосвязей между этими задачами. CRISP-DM предоставляет структурированный подход к интеллектуальному анализу данных, который может помочь обеспечить успех диплому.

Для этого надо понять, какие именно этапы CRISP-DM будут подходить, и подобрать конкретные шаблоны.

Раздел 2. Понимание данных (Data understanding).

Согласно документации CRISP-DM, здесь находятся следующие подразделы и соответствующие им отчеты:

1. Collect Initial Data – Initial Data Collection Report

2. Describe Data – Data Description Report

3. Explore Data – Data Exploration Report

4. Verify Data Quality – Data Quality Report

Были выбраны Initial Data Collection Report (отчёт о сборе данных) и Data Description Report (отчёт об описании данных).

**Initial Data Collection Report.** Данные поступают из различных источников, таких как:

1. Существующие данные. Включают в себя широкий спектр данных, таких как транзакционные данные, данные опросов, веб-журналы и т. д.
2. Купленные данные. Дополнительные данные, например, демографические.
3. Другие данные. Дополнительные данные. Если вышеперечисленные источники не отвечают требованиям, возможно, придется провести опросы или начать дополнительное отслеживание, чтобы дополнить существующие хранилища данных.

Все веб-ресурсы как с готовыми, так и с сырыми данными необходимо указывать. Также нужно отметить, как происходил сбор данных: вручную или с помощью программного кода.

**Data Description Report.** Существует много способов описания данных, но большинство описаний фокусируется на количестве и качестве данных—сколько данных доступно и в каком состоянии они находятся. Ниже перечислены некоторые ключевые характеристики, которые необходимо учитывать при описании данных.

1. Объем данных. Для большинства методов моделирования существуют компромиссы, связанные с размером данных. Большие наборы данных могут создавать более точные модели, но они также могут удлинять время обработки. Надо рассмотреть, возможно ли использование подмножества данных. Делая заметки для окончательного отчета, обязательно включать статистику размеров для всех наборов данных и не забывать учитывать, как количество записей, так и поля (атрибуты) при описании данных.
2. Значимые типы. Данные могут принимать различные форматы, такие как числовой, категориальный (string) или логический (true/false). Обращение внимания на тип значения может предотвратить возникновение проблем во время последующего моделирования.
3. Схема кодирования. Часто значения в базе данных представляют собой представления таких характеристик, как пол или тип продукта. Например, один набор данных может использовать M и F для представления мужского и женского пола, а другой – числовые значения 1 и 2. Стоит обратить внимание на любые конфликтующие схемы в отчете о данных.

Также есть смысл обозначить количественные характеристики данных. Описываются исчислимые показатели в отношении набора данных:

1. Формат данных

2. Способ хранения данных

3. Объем данных (в количестве объектов)

4. Объем данных (в килобайтах)

И качественные характеристики данных.

1. Включают ли в себя данные характеристики, соотносящиеся с целью работы

2. Какие типы данных присутствуют в данных

3. Были ли рассчитаны базовые статистики для ключевых атрибутов (признаков)

4. Существует ли возможность определить важность некоторых признаков:

Раздел 3. Подготовка данных (Data preparation).

Согласно документации CRISP-DM, здесь находятся следующие подразделы и соответствующие им отчеты:

1. Select Data – Rationale for Inclusion / Exclusion

2. Clean Data – Data Cleaning Report

3. Construct Data – Derived Attributes / Generated Records

4. Integrate Data – Merged Data

5. Format Data – Reformatted Data / Dataset / Dataset Description

Были выбраны Data Cleaning Report (отчёт об очистке данных) и Derived Attributes (Производные признаки).

**Data Cleaning Report**. Отчет об использованных способах по очистке данных имеет важное значение для отслеживания изменений в данных. Будущие проекты интеллектуального анализа данных выиграют от того, что детали работы будут легко доступны.

Важно описать следующие Характеристики очистки данных.

1. Описание шума в данных.

2. Алгоритм устранения шума.

3. Потери при устранении шума.

При написании отчета рекомендуется учитывать вопросы:

1. Какие типы шума возникли в данных?
2. Какие подходы использовались для устранения шума? Какие методы оказались успешными?
3. Есть ли какие-нибудь случаи или атрибуты, которые нельзя было бы спасти? Обязательно стоит обратить внимание на данные, исключенные из-за шума.

**Derived Attributes**. Часто бывает так, что нужно построить новые данные. Существует два способа построения новых данных: извлечение признаков (столбцов или характеристик) и генерация записей (строк).

Ниже представлены требования-вопросы к данным для моделирования при извлечении признаков.

1. Ожидает ли алгоритм моделирования определенного типа данных, например, числовых? Если да, то выполнить необходимые преобразования.
2. Нужно ли нормализовать данные перед моделированием?
3. Можно ли построить недостающие признаки с помощью агрегации, усреднения или индукции?
4. Основываясь на фоновых знаниях, существуют ли важные факты, которые могут быть получены из существующих полей?

Также необходимо определить требования к данным до начала построения признаков, выполнить инпутацию пропущенных значений, нормализовать и масштабировать данные и приведение данные в единое признаковое пространство.

Раздел 4. Моделирование (Modelling).

Согласно документации CRISP-DM, здесь находятся следующие подразделы и соответствующие им отчеты:

1. Select Modelling Techniques – Modelling Technique / Modelling Assumptions

2. Generate Test Design – Test Design

3. Build Model – Parameter Setting Model / Model Descriptions

4. Asses Model – Model Assessment, Revised Parameter Settings

Были выбраны Model Descriptions (описание модели) и Modelling Technique (методы моделирования).

**Modelling Technique**. При принятии решения о том, какую модель(и) использовать, стоит подумать, оказывают ли следующие проблемы влияние на выбор:

1. Требует ли модель разбиения данных на тестовые и обучающие наборы?
2. Достаточно ли у данных, чтобы получить достоверные результаты для данной модели?
3. Требует ли модель определенного уровня качества данных? Можно ли соответствовать этому уровню с текущими данными?
4. Являются ли данные подходящим типом для конкретной модели? Если нет, то можно ли выполнить необходимые преобразования с помощью узлов обработки данных?

Что необходимо описать в данном разделе:

1. Описание разделения данных на тренировочные и тестовые.

2. Обоснование количества и качества данных в отношении выбранной модели.

3. Обоснование выбранного типа данных в отношении выбранной модели.

**Model Descriptions**. Изучая результаты работы модели, обязательно нужно делать заметки о своем опыте моделирования. Для каждой модели стоит записать такую информацию, как:

1. Можно ли сделать осмысленные выводы из этой модели?
2. Есть ли новые идеи или необычные паттерны, обнаруженные моделью?
3. Где есть проблемы с исполнением для модели? Насколько разумным было время обработки?
4. Были ли у модели проблемы с качеством данных, такие как большое количество пропущенных значений?
5. Были ли какие-то несоответствия в расчетах, которые следует отметить?

Будет отличной практикой указать описание полученной модели, описание набора гиперпараметров для модели с наилучшими результатами, а также результаты обучения модели.

Источник:

1. <https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf>