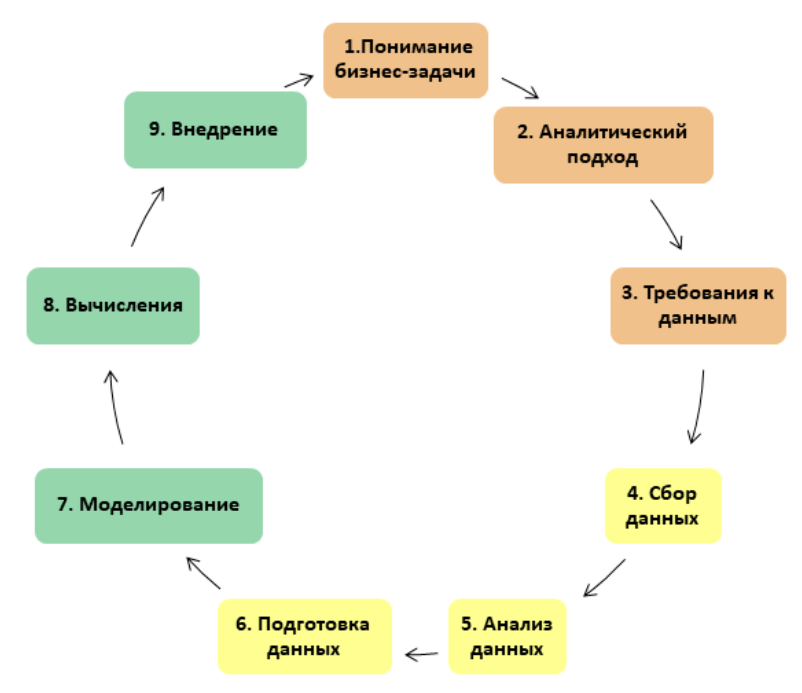
[Предыдущий документ](File%20five.docx)

Даже внедренная модель никогда не может считаться идеальной. Если этапы визуализировать, то получится следующий рисунок:



*Рис. 1 Основные этапы Data Science проекта*

## 

## 2.2 СБОР ДАННЫХ

Нужно отметить, что для объективности испытания нам необходим математический аппарат, так как в ином случае мы не сможем доказать правоту наших утверждений.

При работе с исследованием важен все этапы, так что помимо самих алгоритмов стоит уделить большое внимание сбору информации.

Изучаемы признаки бывают: внутренние и внешние.

**Внутреннии признаки** – определяются личностью студента.

**Внешнии признаки** – определеяются установленной средой, в которой студент живёт.

Так как студенты из разных групп, разных направлений и разных курсов, мы постараемся выделить те признаки, которые универсальны для всех, дабы не строить индивидуальную модель для каждого.

Разработка электронной анкеты. Сбор данных осуществляется в ходе письменного опроса по анкете, включающей N вопросов. Структура анкеты отражает факторы, обозначенные в теоретических вопросах исследования.

Для студентов разных ВУЗ-ов мы построим следующую таблицу:

*Таблица 3. Пример таблицу студентов по ВУЗ-у*

|  |  |  |  |
| --- | --- | --- | --- |
| Направление обучения | Генеральная совокупность | Университет X | Отклонение в % |
| 1. ИВТ | 4,56 | 4,64 | 1,8 |
| 1. Психология | 4,00 | 4,15 | 3,8 |
| 1. Инноватика | 4,41 | 4,59 | 4,1 |

[Предыдущий раздел](#_2.2_СБОР_ДАННЫХ) [Следующий раздел](#_ВЫБОР_ФИЧ)

### **ВЫБОР ФИЧ**

Сразу скажем, что реальные сильно влияющие факторы не могут быть определены сразу, наша задача определить их гарницы. То есть выбрать те фичи, которые с большой вероятностью будут иметь связь с успеваемость. Проверить же это мы сможем только по ходу аналитики данных и фич.

Очевидно, что чем больше факторов мы будем учитывать, тем точнее будут модели, однако избыточные данные нам не нужны, так как они нагрузят наши вычисления, а также могут сильно повлиять не резульаты исследования и онулировать наши старания. Так что наша задача отобрать достаточное количество и необходимое количество признаков.

Основные группы факторов опроса студентов:

* потенциал, уровень знаний, достигнутый к поступлению в вуз;
* мотивация студента к получению высшего образования и причины поступления;
* характер и мотивация студента;
* наличие занятности во внеучебное время;
* влияние среды, родителей;
* преподаватели;
* организация учебного процесса: расписание, работа деканата, библиотеки.

*Таблица 4. Описание переменных полученной анкеты*

|  |  |  |  |
| --- | --- | --- | --- |
| Название признака | Тип данных | Смысл | Пример |
| Ф. И. О. | String | Нужен для идентификации студентов | Иванов Иван Иванович |
| Возраст | Integer | Используется в исследовании, особенно важен при больших отклонениях от моды. | 25 лет |
| Время в часах проводимое за университетской учёбой в неделю | Double | Используется в исследовании, показывать степень затраченных усилий студента | 5.25 часов |
| Время в часах проводимое за развлечениями в неделю | Double | Используется в исследовании, показывает время отдыха студента. | 7.5 часов |
| Время в часах проводимое за хобби в неделю | Double | Используется в исследовании, показывает время для удовлетворения интересов студента. | 1.5 часа |
| Количество секций и кружков, студий | Integer | Показывать активность студента | 5 |
| Укажите количество членов в вашей семье | Integer | Нужен для исследования | 4 |

Пути решения проблем сбора данных:

**1) Параллельные данные**. Нередко оказывается целесообразно разработать две достаточно равноправные методики для измерения заданной социальной характеристики что позволяет установить обоснованность. Когда имеется несколько методови один испольнитель то это так назывемая случай совокупного измерения. Основная задача здесь – установить зависимость двух комплексных оценок с помощью методов регрессионого анализа;

В процессе отработки инструментов измерения в отношении надёжности целесообразна следующая последовательность основных этапов работы:

* предварительный контроль обоснованности методов измерения на стадии проб методики на небольших выборках;
* пилотаж методики и тщательная проверка устойчивости исходных данных и вторичных показателей

усовершенствование методики и доработка всех её деталей и в итоге получение окончательного варианта методики для исследования