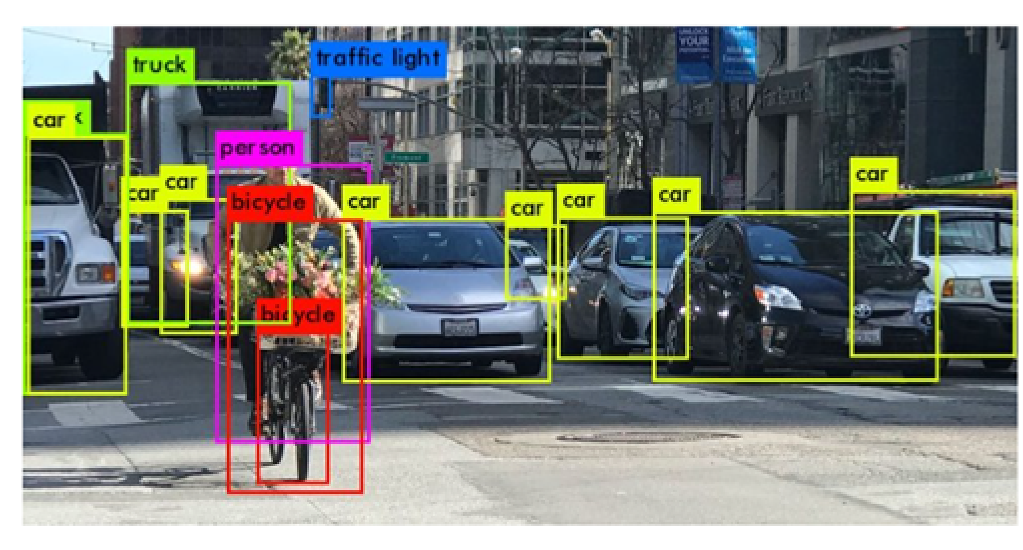
Machine Learning

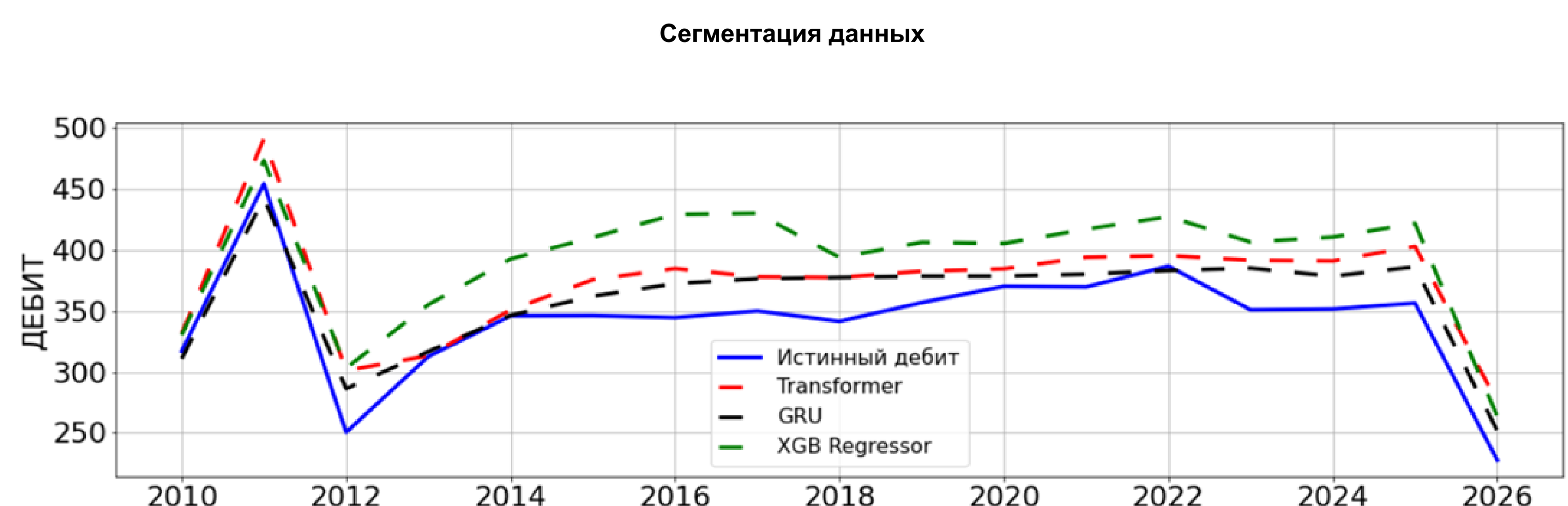
Машинное обучение – область искусственного интеллекта, которая изучает решение задач методом обучения на множестве решений сходных задач.

Решаемые задачи машинного обучения



Детектирование объектов

Сегментация данных



Предсказание нефтедобычи

Методы машинного обучения

Главной особенностью машинного обучения является то, что оно не решает задачи напрямую.

Модель машинного обучения рассматривает набор похожих задач, отличающихся своими входными и выходными данными, на основе этого набора обучается, получая способность решать такие задачи. Соответственно, если с увеличением набора данных возрастает точность решения задач, то мы имеем дело с машинным обучением.

Наиболее популярными методами являются:

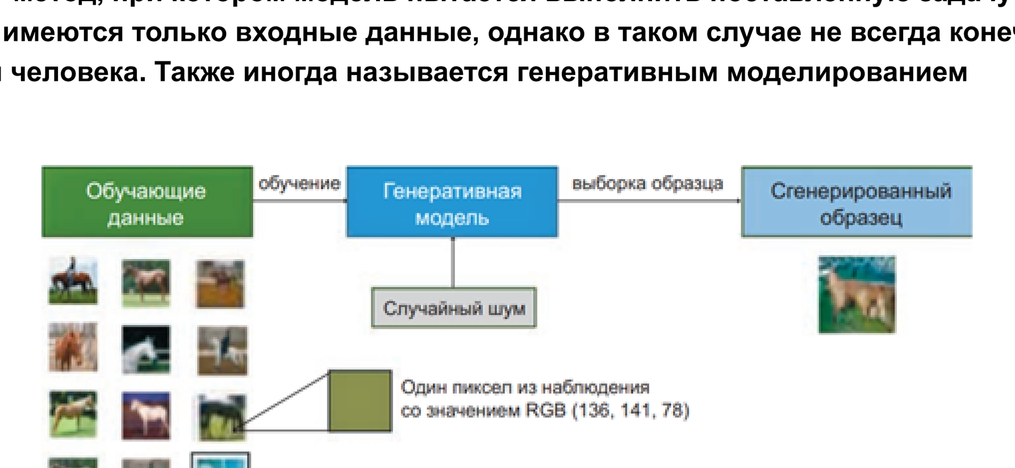
1)Обучение без учителя.

2)Обучение с учителем.

На данном занятии мы остановимся ниъ.

Обучение без учителя

1) Обучение без учителя – метод, при котором модель пытается выполнить поставленную задачу спонтанно, без помощи человека. В таком случае имеются только входные данные, однако в таком случае не всегда конечный результат соответствует ожиданиям человека. Также иногда называется генеративным моделированием



Обучение с учителем

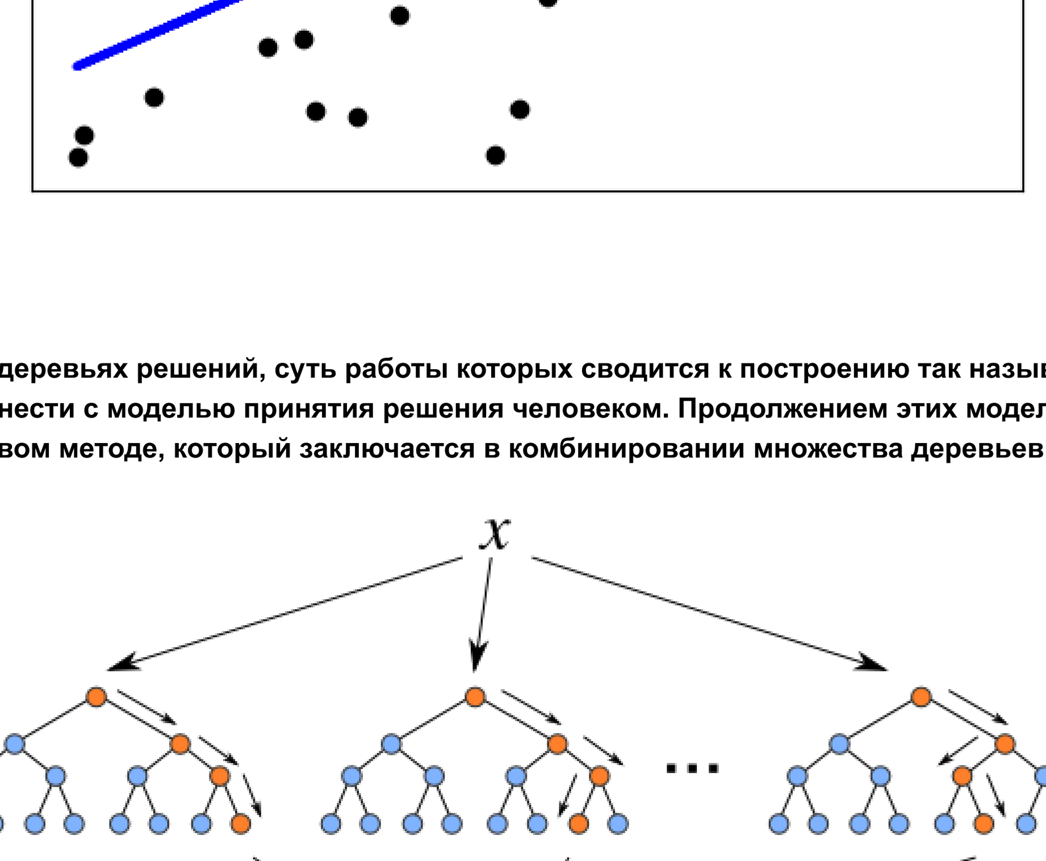
2) Обучение с учителем – метод, при котором человек направляет модель, предоставляя её ей входные и выходные данные, подобно учителю, который на примере объясняет ученику, каков будет результат, если будет произведены определенные действия. Также иногда обучение с учителем называют дискриминативным моделированием.

Стоит отметить, что обучение с учителем является наиболее распространенным методом машинного обучения, так как обучение без учителя может быть используемо только в крайне ограниченном количестве задач (например, кластеризация данных).

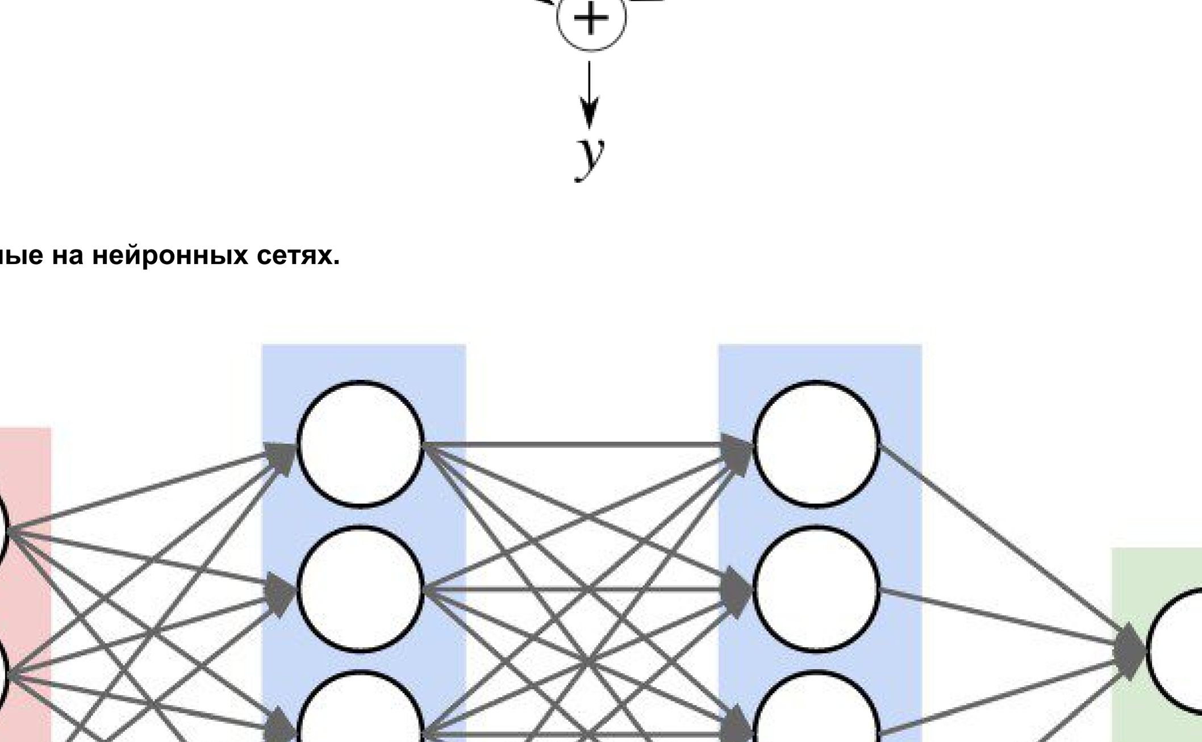
Методом обучения с учителем решают такие классические задачи, как регрессия и классификация. Цель задачи регрессии получить число или ряд чисел, цель классификации - определение класса объекта из списка заданных. В обоих случаях обучение также происходит на базе входных данных, имеющих свои определенные параметры, и на основании выходных данных (чисел для регрессии и классов для классификации).

При решении данных задач при помощи машинного обучения можно также использовать большое количество моделей, таких как:

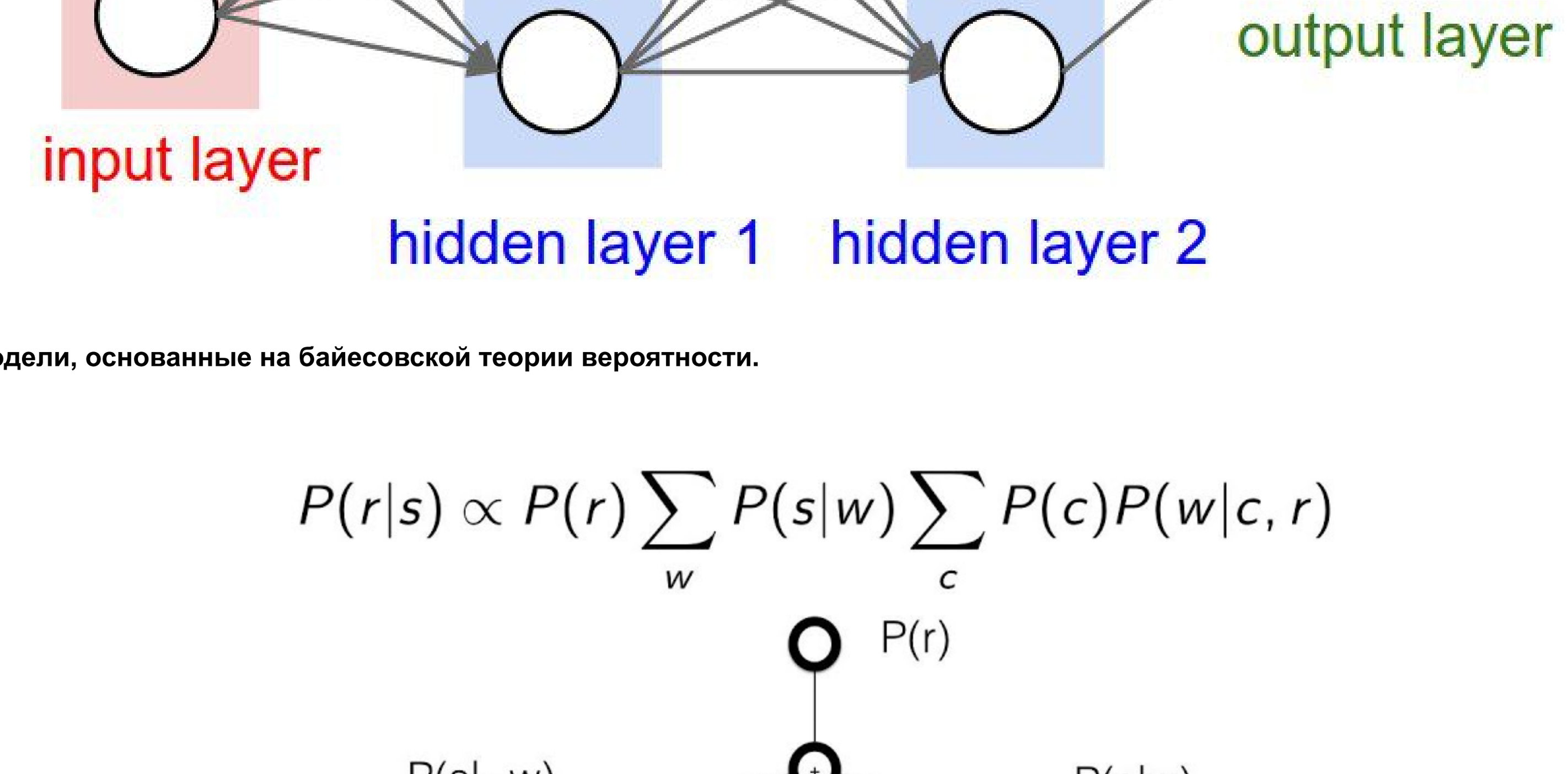
•Линейные модели, цель которых построить разделяющую (для классификации) или аппроксимирующую (для регрессии) гиперплоскость.



•Модели, основанные на деревьях решений, суть работы которых сводится к построению так называемого дерева решений, что можно соотнести с моделью принятия решения человеком. Продолжением этих моделей являются модели, основанные на ансамблевом методе, который заключается в комбинировании множества деревьев.

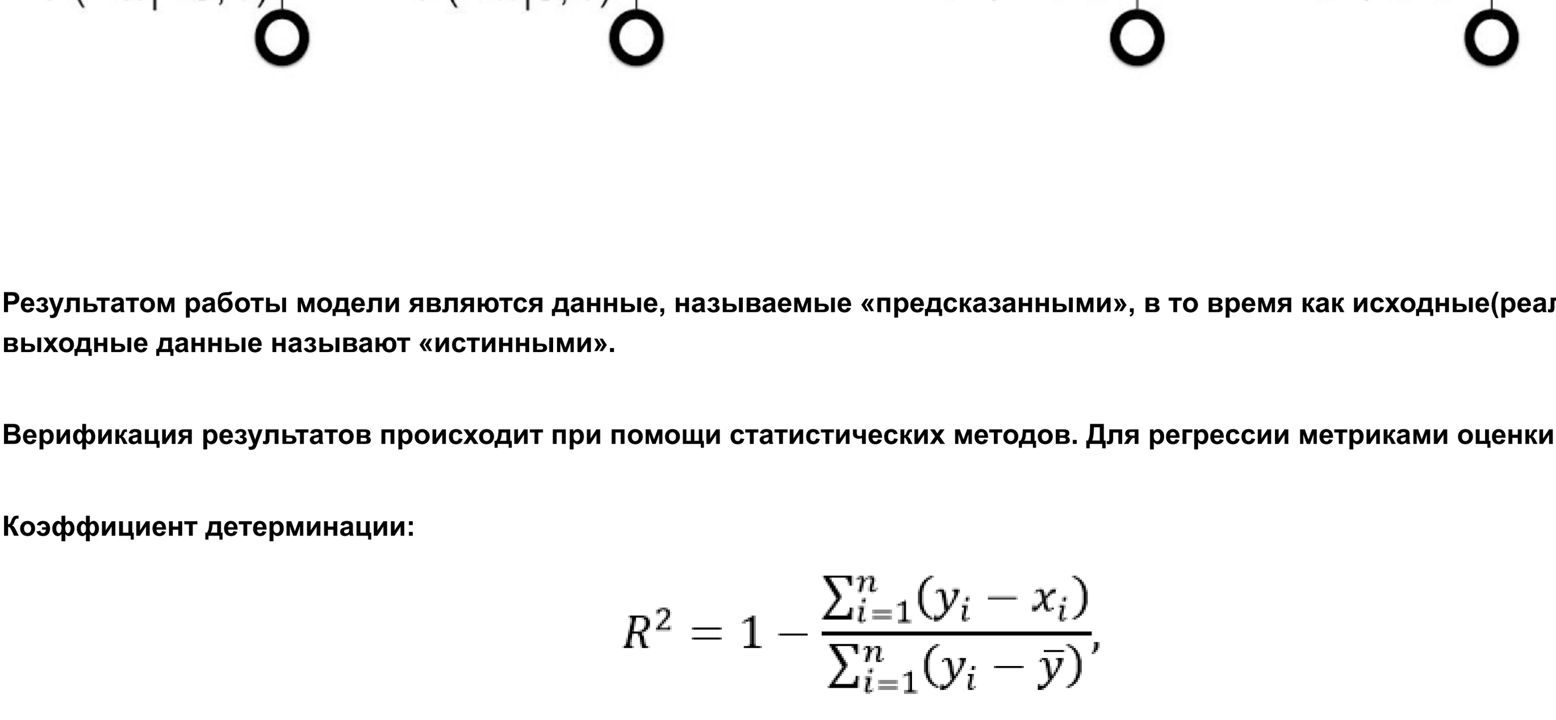


•Модели, основанные на нейронных сетях.



•Модели, основанные на байесовской теории вероятности.

$$P(r|s) \propto P(r) \sum_{w} P(s|w) \sum_{c} P(c) P(w|c, r)$$



Результатом работы модели являются данные, называемые «предсказанными», в то время как исходные(реальные) выходные данные называют «истинными».

Верификация результатов происходит при помощи статистических методов. Для регрессии метриками оценки служат:

Коэффициент детерминации:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)}{\sum_{i=1}^{n}(y_i - \bar{y})},$$

Средняя абсолютная ошибка:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|.$$

Средняя абсолютная ошибка в процентах:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{x_i}\right|.$$

Для классификации:

Кросс-энтропия

$$P = \prod_{i=1}^{n} p(y_i|x_i),$$

$$-ln(P) = \sum_{i=1}^{n} ln(p(y_i|x_i)),$$

где $p(y_i|x_i)$ - вероятность принадлежности объекта x_i к классу y_i.

При оценке результатов перечисленными метриками стоит быть осторожными. При различных задачах те или иные метрики могут быть не репрезентативными, а неверная оценка может привести к неверным выводам – например, о плохой работе машинного обучения применительно к выбранной задаче.

Также огромную роль в машинном обучении играет база данных и её качество. Во-первых, данных должно быть достаточно для применения машинного обучения (размер варьируется от задачи к задаче, но как правило, для корректной работы количество так называемых «экспериментов» должно исчисляться тысячами, десятками тысяч и более). Во-вторых, данные должны быть проанализированы и корректно обработаны. Для классификации данных нередко их необходимо размечать.

Стоит отметить, что при обучении модель делает это не на всех данных, а только лишь на её части. Как правило база данных делится случайным образом в соотношении 80%/20%, где 80% называют обучающей выборкой, а 20% - тестовой. Однако всегда возникает риск того, что среди 80% случайным образом окажется много похожих друг на друга экспериментов, которые не встречаются в тестовой выборке, либо наоборот. Оценка результатов в таком случае может привести к ложной иллюзии о том, что модель работает слишком хорошо, а может и наоборот, слишком плохо. Во избежание этого при обучении база данных может разбиваться на части случайным образом несколько раз, однако надёжнее использовать кросс-валидацию. Это разбиение базы данных на, к примеру, 10 частей (стандартное разбиение), то есть база данных делится в соотношении 90%/10% 10 раз, при этом на каждом разбиении тестовая выборка не включает в себя эксперименты из других тестовых выборок. Таким образом мы гарантируем, что наша модель будет обучаться и валидироваться на всех экспериментах, входящих в базу данных.

Таким образом, когда мы хотим решить какую-либо прикладную задачу методом машинного обучения, необходимо:

1) Четко сформулировать саму задачу, в том числе понимать, что должно быть результатом её решения.

2) Выбрать метод решения (с учителем, без него или другие).

3) Выбрать способ решения задачи (регрессия, классификация, кластеризация).

4) Выбрать модель для решения этой задачи.

5) Подготовить базу данных для работы с моделью.

6) Обучить модель.

7) Верифицировать результаты.

8) Если качество работы низкое, предлагается увеличить базу данных, найти и убрать аномальные данные, которые мешают работе модели, либо выбрать другую модель.