



# MACHINE LEARNING



# ПЛАН МЕРОПРИЯТИЯ

- 1 - ОПРЕДЕЛЕНИЕ
- 2 - ДАННЫЕ И ПРИЗНАКИ ОБЪЕКТОВ
- 3 - МЕТОДЫ ОБУЧЕНИЯ И МОДЕЛИ
- 4 - ПРОБЛЕМЫ, СВЯЗАННЫЕ С ОБУЧЕНИЕМ
- 5 - ЗОЛОТЫЕ ПРАВИЛА ОБУЧЕНИЯ МОДЕЛЕЙ
- 6 - LIVE CODE ОБУЧЕНИЕ ML



**Машинное обучение**<sup>[1]</sup> (англ. *machine learning*, ML) — класс методов **искусственного интеллекта**, характерной чертой которых является обучение за счёт решений множества сходных задач.

Целью машинного обучения является частичная или полная автоматизация решения сложных профессиональных задач в самых разных областях человеческой деятельности.



## Artificial intelligence

Natural language processing

Visual perception

Automatic programming

Intelligent robot

Knowledge representation

Automatic reasoning

## Machine learning

Linear/Logistic regression

k-Means

Support vector machine

Principal component analysis

k-Nearest neighbor

Random forest

Decision trees

## Neural Networks

Boltzmann neural networks

MLP

## Deep Learning

CNN      RNN

GAN      DBN

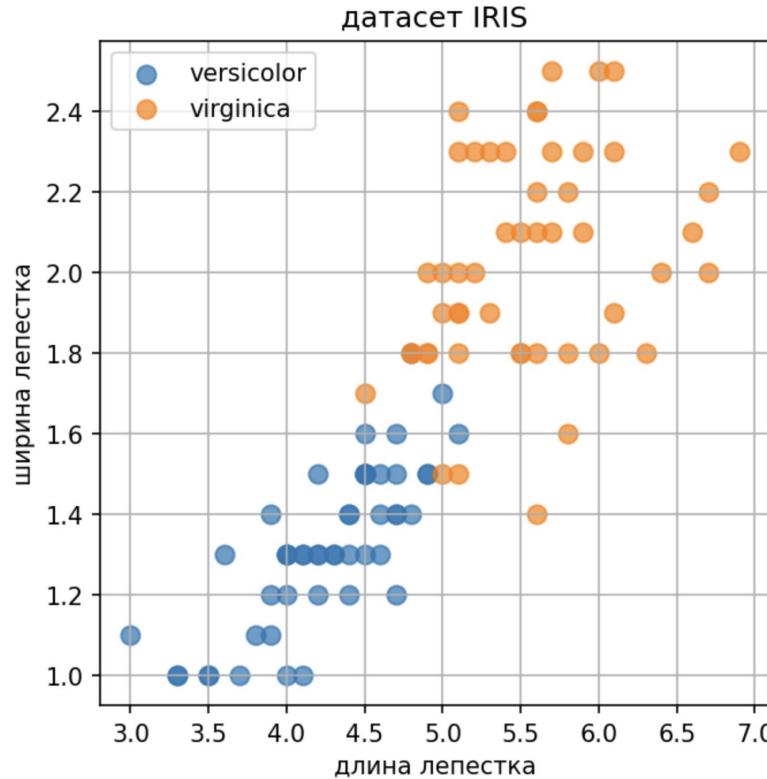


# ДАННЫЕ - ОСНОВА ДЛЯ ML

X - ОБУЧАЮЩАЯ ВЫБОРКА  
ширина и длина лепестка

Y - ИЗВЕСТНЫЕ ОТВЕТЫ  
соответственный вид растения

ЗАДАЧА ML - найти алгоритм для  
выявления зависимости в данных





## ОБЪЕКТЫ И ПРИЗНАКИ

Объект — это элемент выборки, характеризуемый измеряемыми или вычисляемыми признаками

Типы признаков -  
числовые / категориальные / бинарные  
порядковые

складывается матрица -  
объект-признаки-

petal_length	petal_width	species
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa

X

Y



## МЕТОДЫ ОБУЧЕНИЯ В МЛ

# Classical Machine Learning

Task Driven



## Supervised Learning

( Pre Categorized Data )

Data Driven



## Unsupervised Learning

( Unlabelled Data )



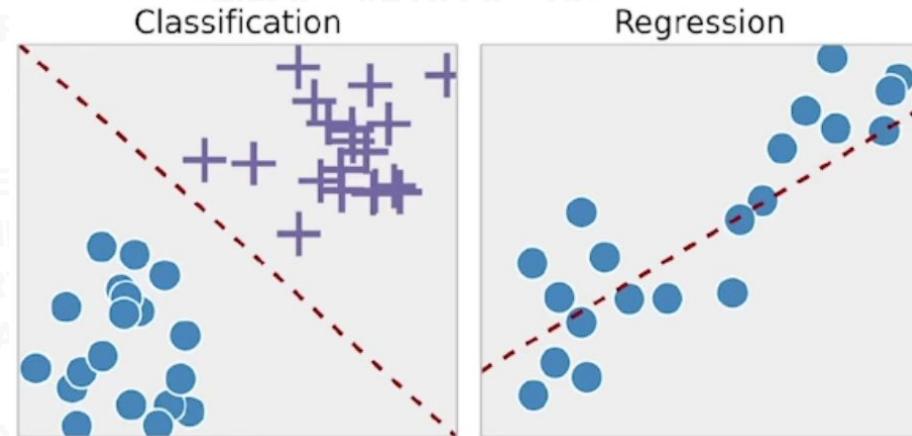
# ДВА ОСНОВНЫХ ТИПА ЗАДАЧ В ОБУЧЕНИИ С УЧИТЕЛЕМ

Задачи классификации (classification):

- классификация на 2 класса.
- классификация на N классов

Задачи восстановления регрессии

- восстановление зависимостей по эмпирическим данным





# ЗАДАЧА КЛАССИФИКАЦИИ

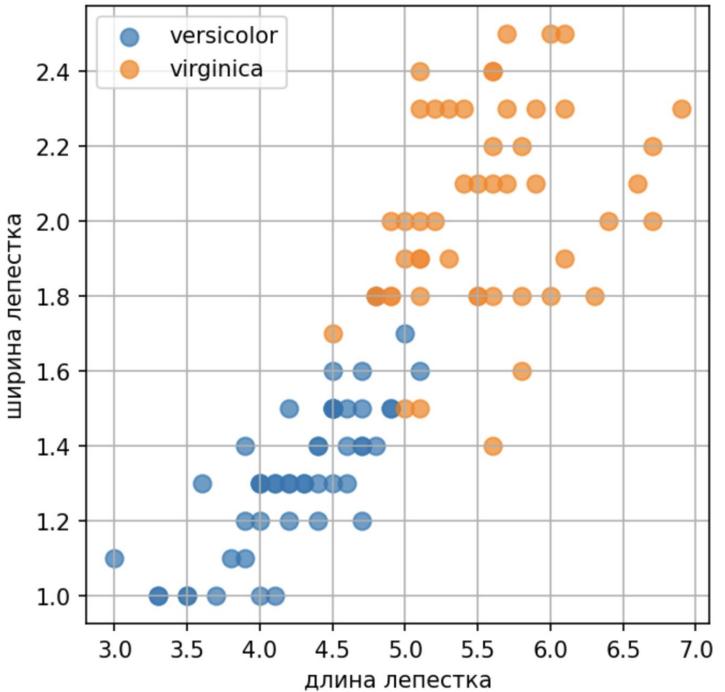
Качественные и количественные признаки Целевая переменная

	encounter_id	gender	age	weight	time_in_hospital	num_lab_procedures	diag_1	diag_2	diag_3	diabetesMed	readmitted	
0	2278392	Female	[0-10)	?	1		41	250.83	?	?	No	0
1	149190	Female	[10-20)	?	3		59	276	250.01	255	Yes	0
2	64410	Female	[20-30)	?	2		11	648	250	V27	Yes	0
3	500364	Male	[30-40)	?	2		44	8	250.43	403	Yes	0
4	16680	Male	[40-50)	?	1		51	197	157	250	Yes	0

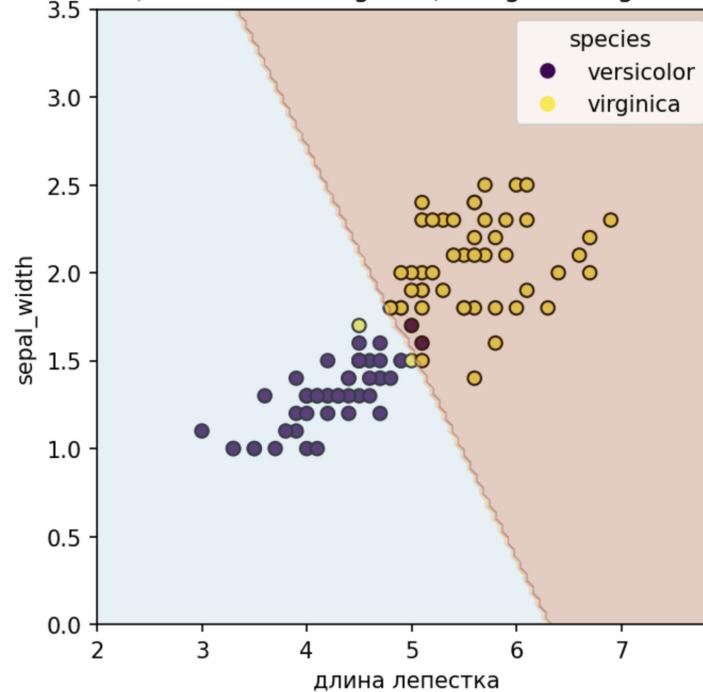


# ПРИМЕР ЗАДАЧА КЛАССИФИКАЦИИ

## датасет IRIS



Iris (versicolor vs virginica) - Logistic Regression

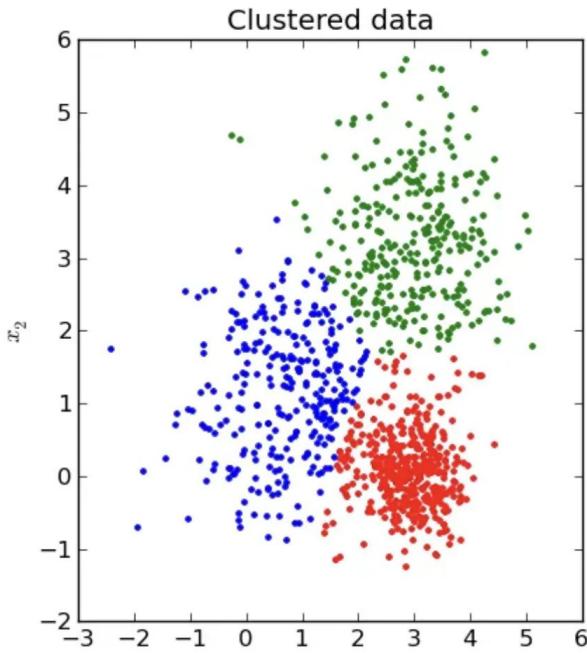
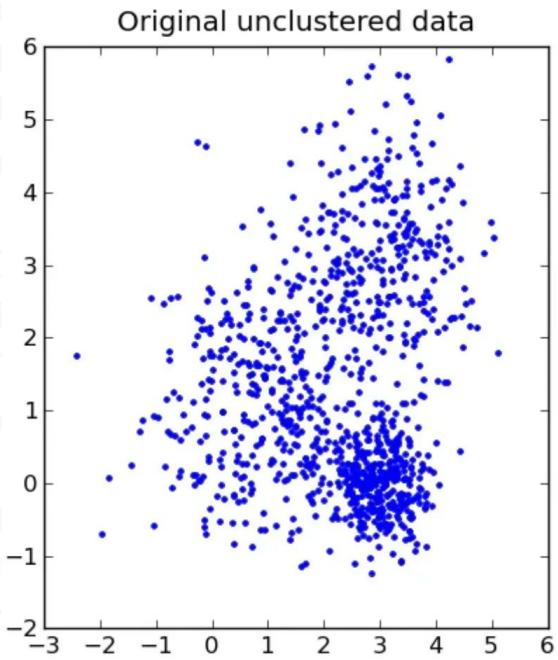




# ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

## Неразмеченные данные

- Сегментация клиентов
- Сегментация пациентов по фармакологическому ответу
- Кластеризация товаров в интернет-магазине



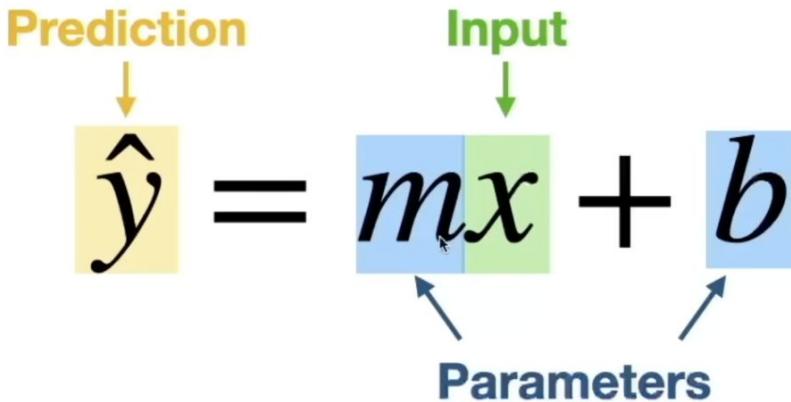


# ЧТО ТАКОЕ МОДЕЛЬ?

Модель (predictive model) -  
параметрическое семейство функций

Выявляет зависимость  $Y$  от  $X$  с учетом  
настраиваемых параметров и  
собственной архитектуры

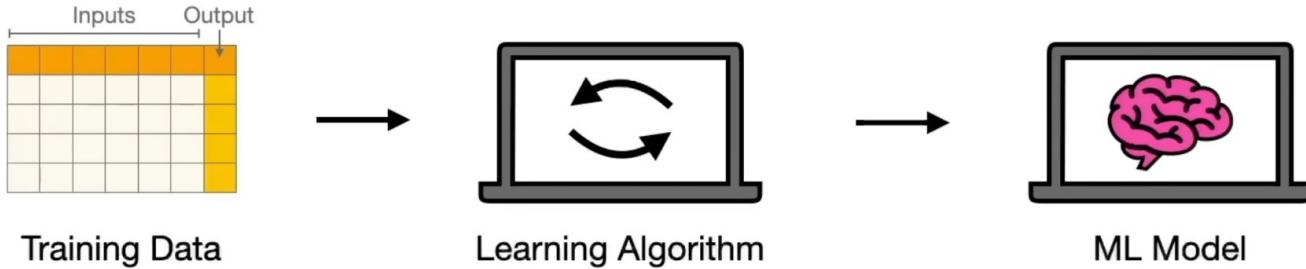
Данные используются для  
оптимизации параметров модели



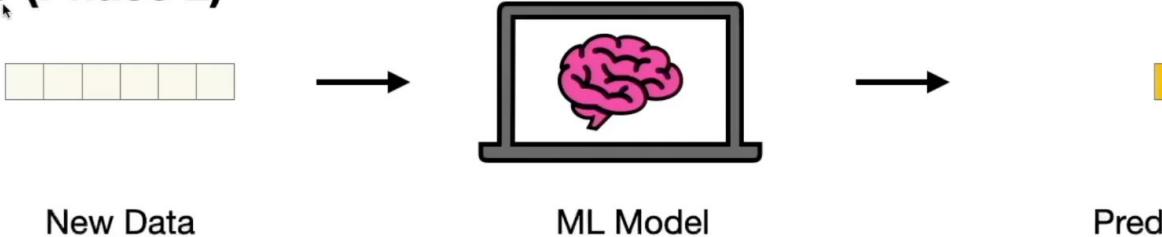


# ЭТАПЫ ОБУЧЕНИЯ И ПРИМЕНЕНИЯ МОДЕЛИ

## Training (Phase 1)



## Inference (Phase 2)





# ФАЗА 1 - ОБУЧЕНИЕ

Подбор параметров модели для минимизации ошибки моделирования (задача оптимизации)

Функция потерь (loss function) — величина ошибки алгоритма

Для задач классификации loss = неверный класс

Для задач регрессии loss = абсолютная и квадратичная ошибки

Методом минимизации эмпирического риска подбираются лучшие параметры

$$L(m, b) = \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \left( m \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right) \right\|$$

**Loss Function**

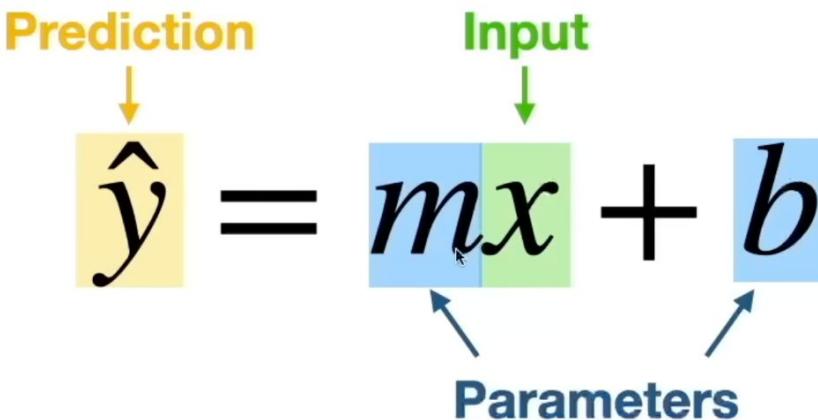
↑                              ↑  
**Actual Values**            **Predicted Values**



## ФАЗА 2 - ПРЕДСКАЗАНИЯ (INFERENCE)

Обученная модель используется  
для предсказаний  
на вход X - на выход Y

X - длина и ширина лепестка  
m & b - параметры модели  
Y - вид цветка





# МОДЕЛИ / ФУНКЦИИ ПОТЕРЬ / ТИП ЗАДАЧИ

Name	Description	Loss Function	Type
Linear Regression	Predicts continuous output by fitting a linear relationship between inputs and output	Mean Squared Error (MSE)	Regression
Logistic Regression	Models the probability of a binary outcome using a logistic (sigmoid) function	Binary Cross-Entropy (Log Loss)	Classification
Decision Tree	Splits data into branches based on feature values to make predictions	Impurity measures (e.g., Gini, Entropy, MSE)	Both
Random Forest	Ensemble of decision trees averaged (regression) or voted (classification)	Same as Decision Tree	Both
XGBoost	Gradient boosting framework that builds trees sequentially to correct prior errors	Customizable; often Log Loss or MSE	Both
SVM (Support Vector Machine)	Finds the optimal hyperplane that separates classes or fits data	Hinge loss (classification), Epsilon-insensitive loss (regression)	Both



# АНАЛОГИЯ ИЗ МЕДИЦИНЫ

## Шкала Апгар оценки состояния новорожденного

Признаки	0 баллов	1 балл	2 балла
Пульс	Отсутствует	Менее 100 уд./мин	Более 100 уд./мин
Дыхание	Отсутствует	Медленное, нерегулярное	Хорошее, крик
Мышечный тонус	Слабый	Сгибает ручки и ножки	Активно двигается
Рефлексы (реакция на катетер в носу)	Отсутствует	Гrimасы	Чихает, кашляет, отталкивает
Цвет кожи	Синюшный, бледный	Нормальный, но синюшные ручки и ножки	Нормальный по всему телу

**10-7 баллов** состояние отличное или хорошее

**6-4 балла** состояние удовлетворительное

**3-0 баллов** состояние неудовлетворительное (критическое)



# ПРОБЛЕМЫ, СВЯЗАННЫЕ С ОБУЧЕНИЕМ

Найдём ли мы «закон природы» или переобучимся?

- переобученная модель идеально подогнана под обучающие данные

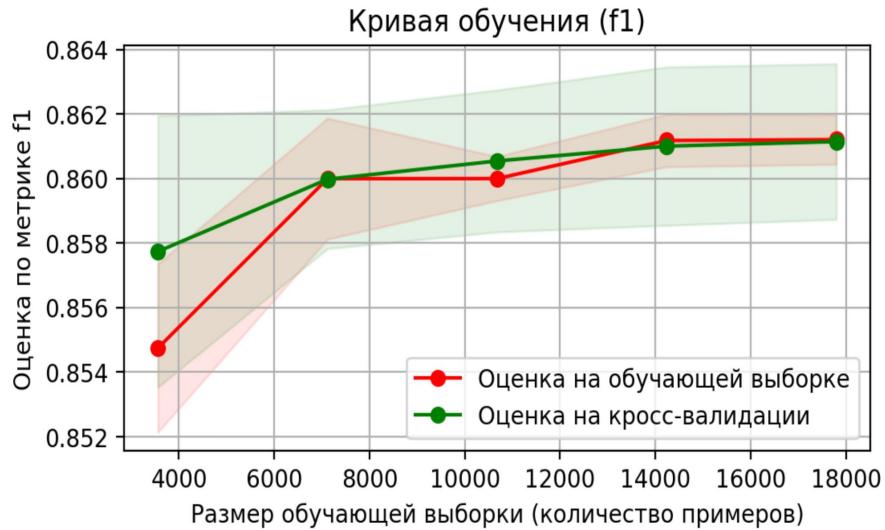
Будет ли модель эффективна на всей генеральной совокупности?

- необходимо определить точность модели после обучения

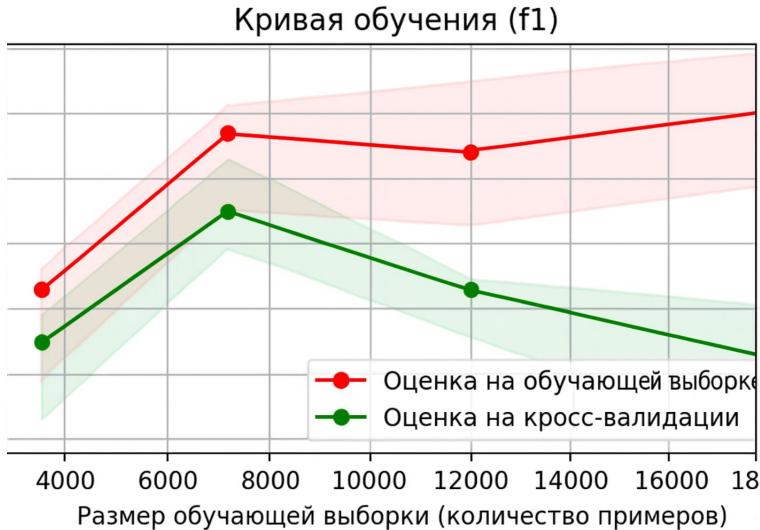


# ПРОБЛЕМЫ, СВЯЗАННЫЕ С ОБУЧЕНИЕМ

## нормальное обучение



## переобучение





# ЗОЛОТЫЕ ПРАВИЛА ОБУЧЕНИЯ МОДЕЛЕЙ

- разделение всей выборки данных на обучающую и валидационную `train_test_split` (метод Hold-Out)
- процедуры препроцессинга применяем по-отдельности к обучающей и тестовой выборкам - избегаем утечки данных
- Использование кросс-валидации — так получается более надежная и стабильная оценка
- Оценка качества модели должна быть адекватной задаче (подбор метрик)

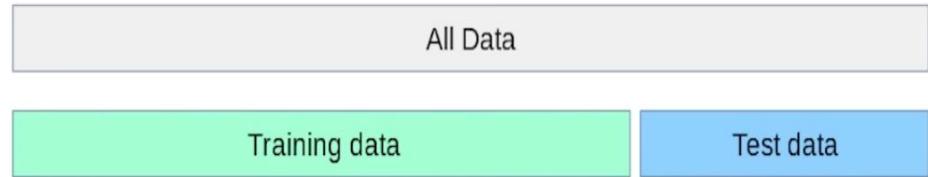


# РАБОТА С ВЫБОРКАМИ

метод Hold-Out `train_test_split`

для обучения - `train`

для финальной оценки  
модели - `test`



позволяет измерить метрики качества модели  
без необходимости в новых данных

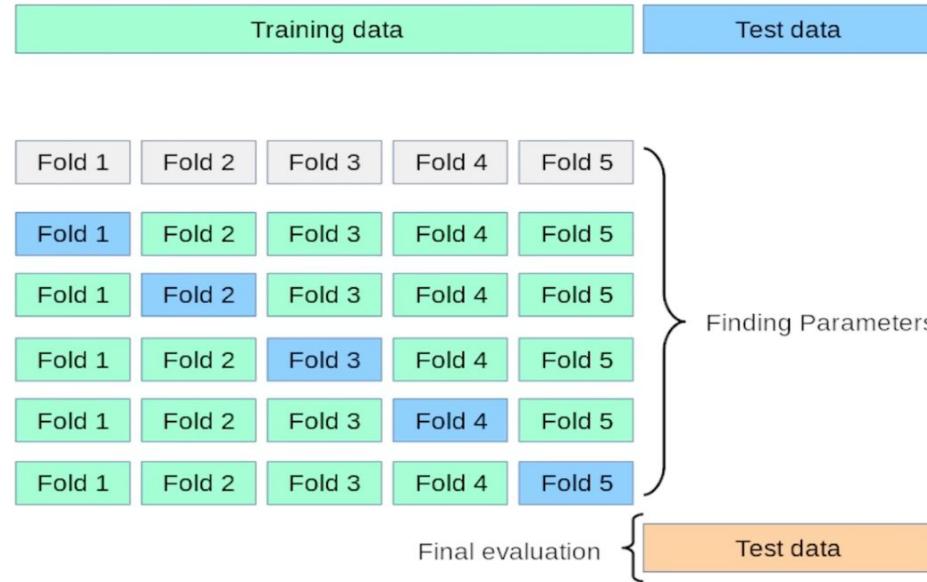


# РАБОТА С ВЫБОРКАМИ

метод K-Fold кросс-валидация  
для обучения - train

для оптимизации  
гиперпараметров - CV

для финальной оценки  
качества - test



## CROSS VALIDATION

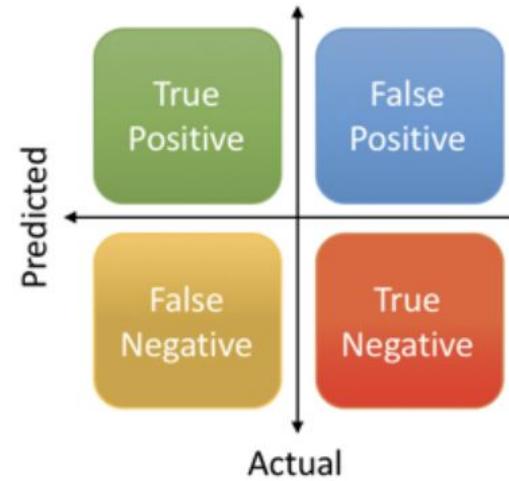


# МЕТРИКИ ДЛЯ КЛАССИФИКАЦИИ

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



Precision (специфичность) — доля действительно активных среди предсказанных

Recall (чувствительность) — доля верно распознанных активных соединений

Accuracy — общая точность предсказания



# МЕТРИКИ ДЛЯ РЕГРЕССИИ

- 1)  $R^2$  коэффициент детерминации который показывает какая доля дисперсии целевых значений объясняется моделью  $R^2 = 1$   $R^2 = 0$
- 2)  $MSE$  Mean Squared Error показывает насколько в среднем прогнозы модели отклоняются от реальных значений в квадрате - очень чувствительна к выбросам  $MSE(y, \hat{y})$
- 3)  $RMSE$  Root Mean Squared Error  $RMSE(y, \hat{y})$



# СТЕК MACHINE LEARNING



ANACONDA.

DATA



seaborn

ФРЕЙМВОРКИ



scikit  
learn TensorFlow

PyTorch



## ПОЛЕЗНЫЕ РЕСУРСЫ

- Kaggle платформа для организации конкурсов и соревнования
- Хендбук Машинное обучение от Яндекс Образование ШАД
- Ускоренный курс ML от Google
- VS Code + всевозможные расширения
- Google Colab запуск Jupyter в облаке

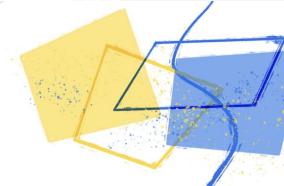


Машинное обучение

Освойте современные методы ML: от классического обучения до глубинного обучения и нейросетей, генеративных моделей и устройства рекомендательных систем.

Нужны знания Python и основы математики

[перейти к изучению](#)



Introduction to  
Machine Learning

