



Московский институт электроники и  
математики им. А.Н. Тихонова

Кафедра информационной  
безопасности киберфизических  
систем

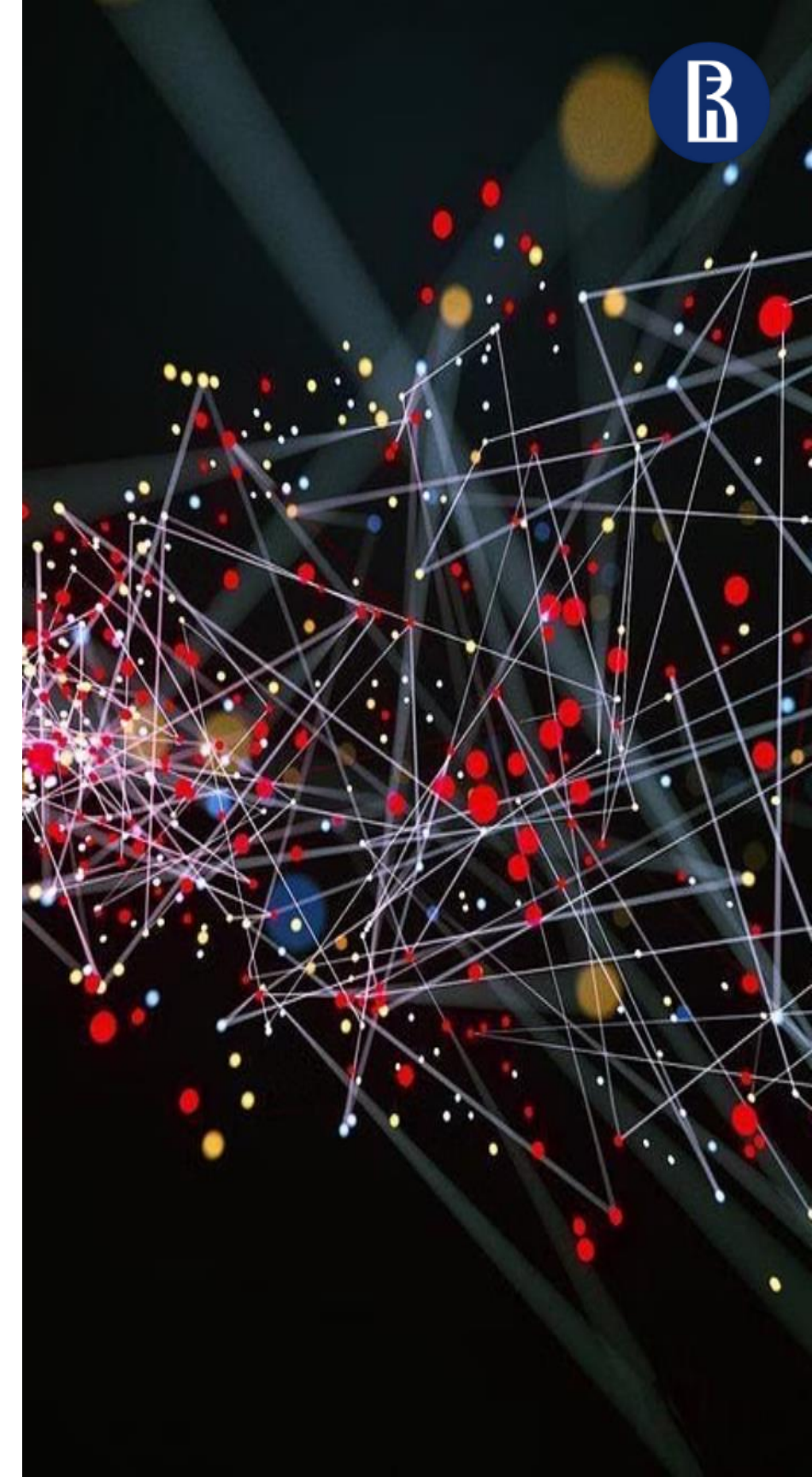
Москва 2025

# Изучение возможностей ML-моделей

# Введение в ML-модели в RapidMiner

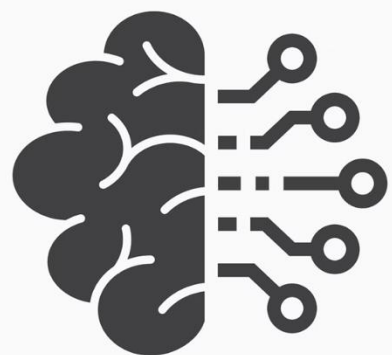
RapidMiner позволяет изучать модели машинного обучения без программирования.

Визуальный интерфейс упрощает использование алгоритмов (Decision Tree, Logistic Regression, k-NN, Random Forest, SVM) и оценку качества через Cross Validation.





# Зачем нужны разные ML-модели?



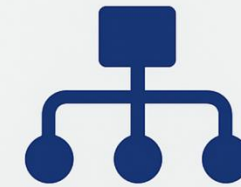
MACHINE  
LEARNING

Использование нескольких моделей важно для:

- Подбора лучшего подхода к конкретной задаче
- Учёта разных типов данных
- Сравнения эффективности и точности различных методов обучения

# Обзор используемых алгоритмов

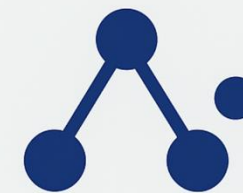
- Decision Tree (дерево решений)
- Logistic Regression (логистическая регрессия)
- k-NN (метод ближайших соседей)
- Random Forest (случайный лес)
- SVM (машина опорных векторов)



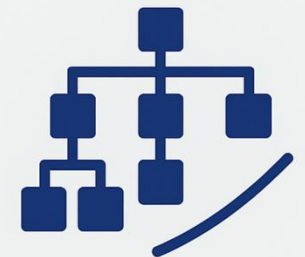
Decision Tree



Logistic Regression



k-NN



Random Forest



# Датасет «Wine Quality»

Набор данных содержит химические характеристики вин (кислотность, pH, алкоголь и др.) и оценку качества.

Задача: классификация вина на «хорошее» и «не очень хорошее».

<new process> – Altair AI Studio Educational 2024.1.0 @ MacBook-Air-Nasta-3.local

Views: Design Results Turbo Prep Auto Model Interactive Analysis Find data, operators...etc All Studio

Result History ExampleSet (//Local Repository/winequality-red)

Open in Turbo Prep Auto Model Interactive Analysis Filter (1,599 / 1,599 examples): all

Row No.	fixed acidity	volatile aci...	citric acid	residual su...	chlorides	free sulfur ...	total sulfur...	density	pH	sulphates	alcohol	q
1	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5
2	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800	5
3	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.800	5
4	11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	9.800	6
5	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5
6	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400	5
7	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400	5
8	7.300	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470	10	7
9	7.800	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570	9.500	7
10	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	5
11	6.700	0.580	0.080	1.800	0.097	15	65	0.996	3.280	0.540	9.200	5
12	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	5
13	5.600	0.615	0	1.600	0.089	16	59	0.994	3.580	0.520	9.900	5
14	7.800	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560	9.100	5
15	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880	9.200	5

ExampleSet (1,599 examples, 0 special attributes, 12 regular attributes)



# Почему классификация вина важна?

Экспертная оценка вин затратна по времени и средствам.

Автоматическая классификация позволяет сократить расходы и ускорить процесс контроля качества, используя только химические данные.





# Основные этапы анализа в RapidMiner

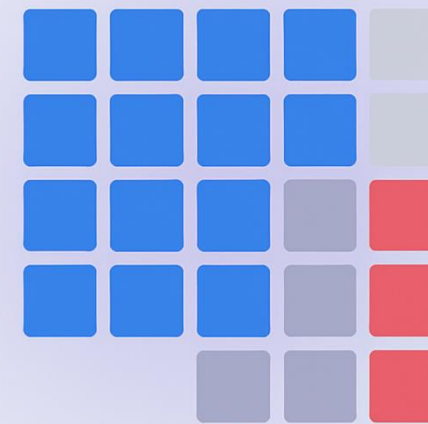


- Загрузка данных (Read CSV)
- Предобработка и создание признаков (Generate Attributes)
- Установка ролей данных (Set Role)
- Обучение моделей
- Оценка качества (Cross Validation)

# Что такое Cross Validation?

Метод оценки качества модели, когда данные делятся на несколько блоков.

Модель обучается и тестируется на разных наборах, обеспечивая устойчивую и объективную оценку точности.



Training sets

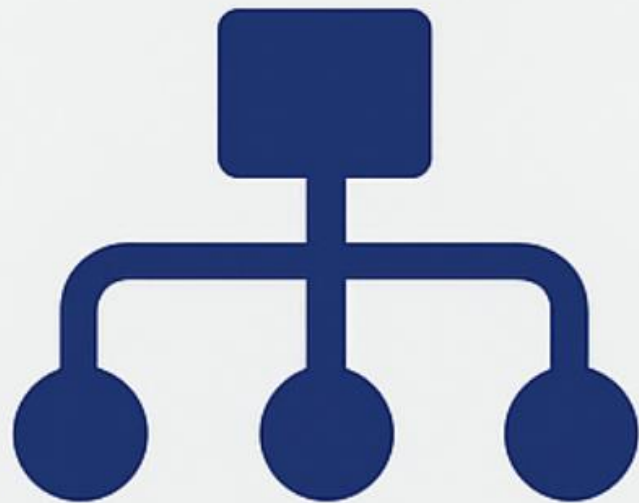
Test set







# Decision Tree: принцип работы



Модель разбивает данные на группы вопросами (« $\text{pH} > 3.2?$ », « $\text{alcohol} > 10?$ »).

Конечный лист определяет класс вина. Логика простая, легко визуализируется и интерпретируется.

# Decision Tree:

## преимущества и недостатки

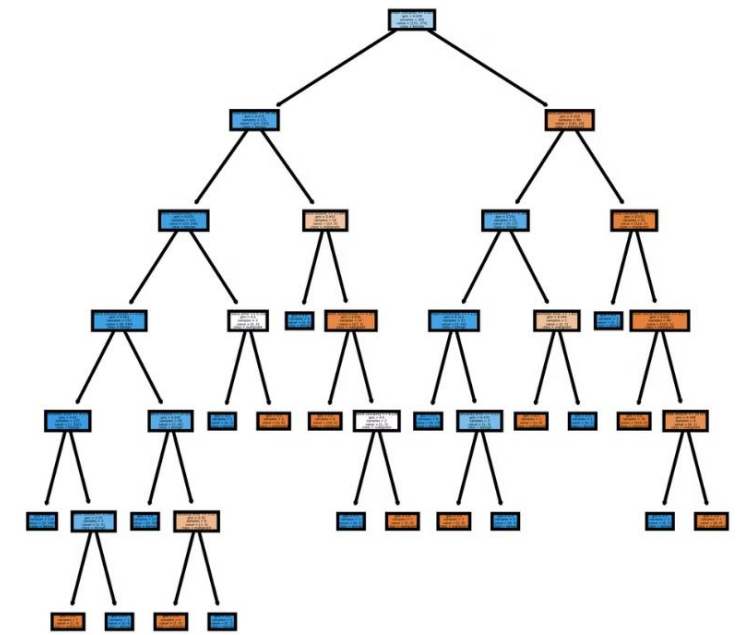


### Преимущества:

- Простота интерпретации
- Наглядность результатов

### Недостатки:

- Склонность к переобучению
- Чувствительность к шумам в данных





# Logistic Regression: суть метода



Логистическая регрессия оценивает вероятность принадлежности объекта к классу, используя линейную комбинацию признаков.

Это простой и эффективный подход, часто используемый в базовых задачах классификации.

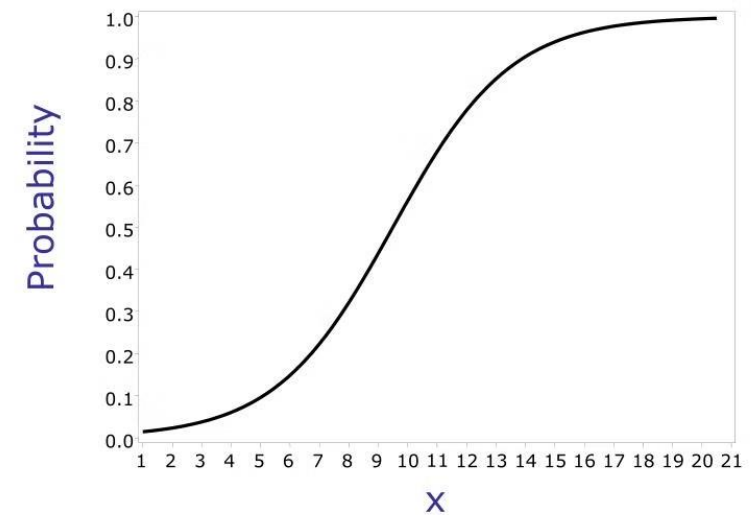
# Logistic Regression: интерпретация



Коэффициенты логистической регрессии показывают влияние каждого признака на вероятность принадлежности к классу.

Положительные коэффициенты увеличивают вероятность, отрицательные — уменьшают.

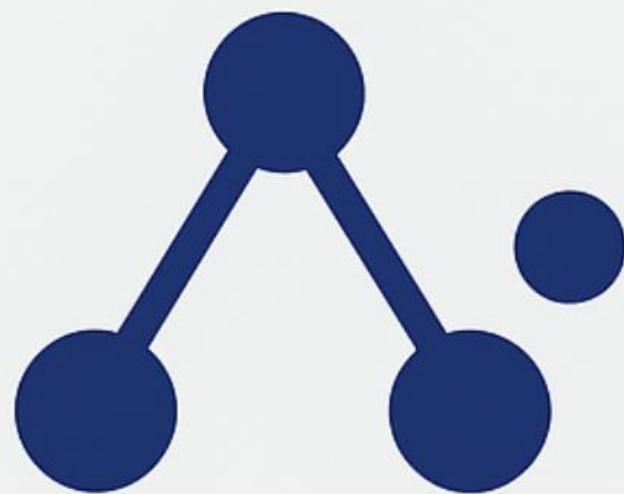
Logistic Regression Curve





## **k-NN: принцип классификации**

Метод k ближайших соседей относит объект к тому классу, который преобладает среди ближайших k объектов.



Требует нормализации признаков для точности работы.

# Особенности настройки k-NN

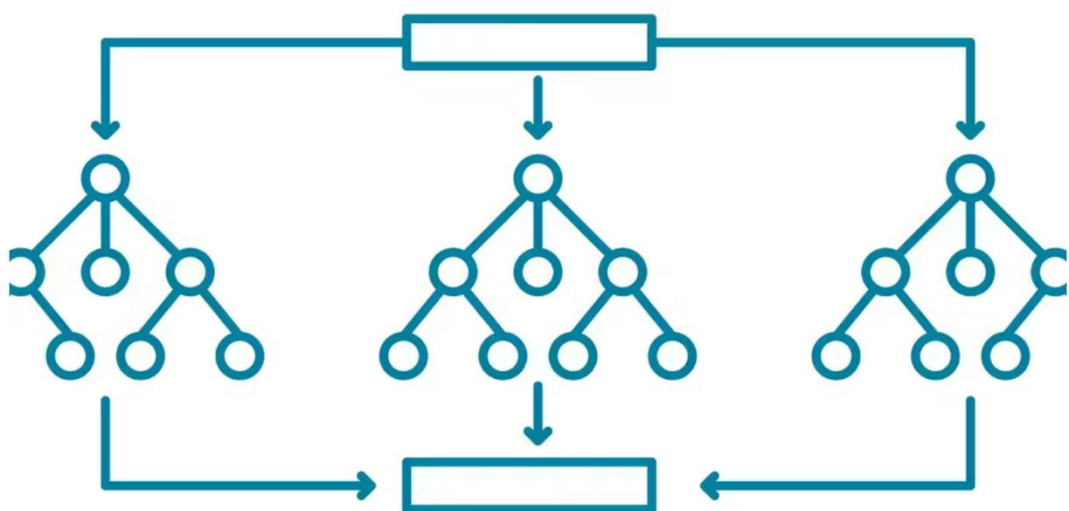
- Число соседей ( $k$ ) критически влияет на качество.
- Метрика расстояний (евклидово, манхэттенское) изменяет чувствительность метода.
- Нормализация данных обязательна для качественного результата.







# Random Forest: ансамбль деревьев



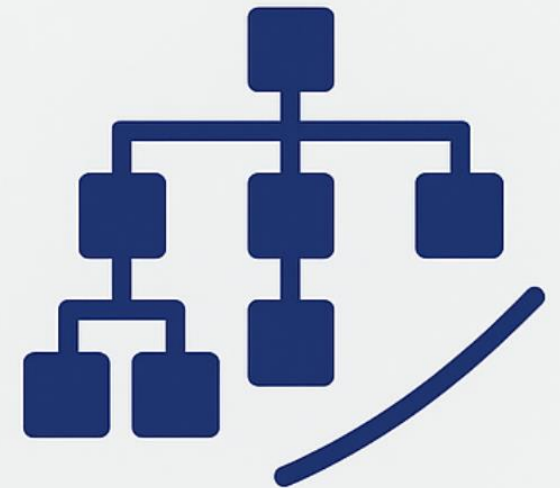
Random Forest обучает множество деревьев решений на случайных подвыборках данных и признаков, затем агрегирует результаты.

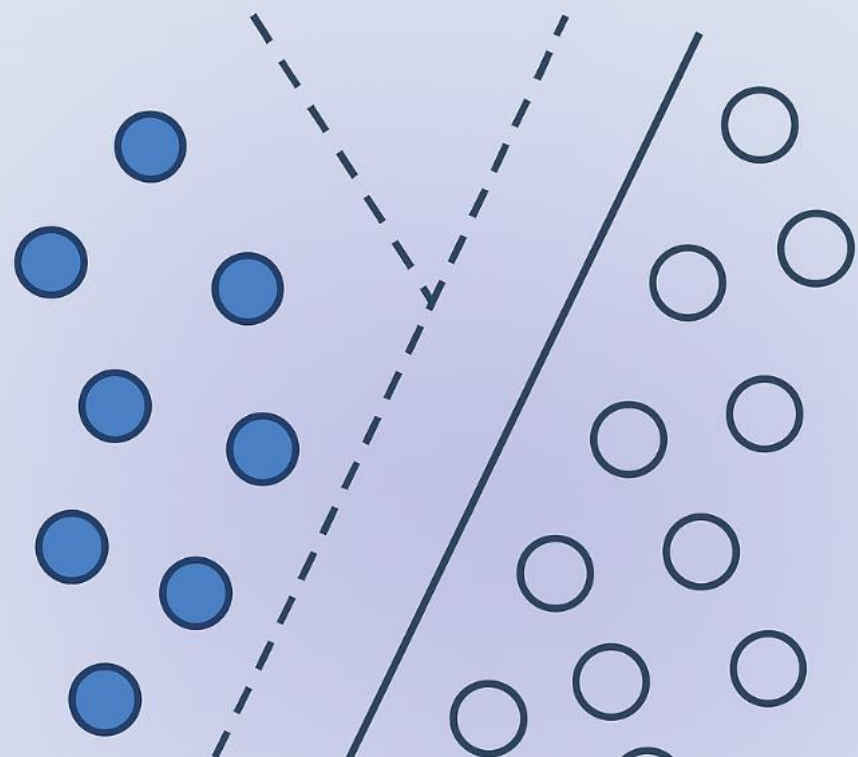
Это уменьшает риск переобучения и повышает стабильность модели.

# Преимущества и сложность Random Forest

Преимущества: высокая точность, устойчивость к шумам.

Недостатки: сложности в интерпретации из-за большого числа деревьев и увеличенные требования к вычислительным ресурсам.





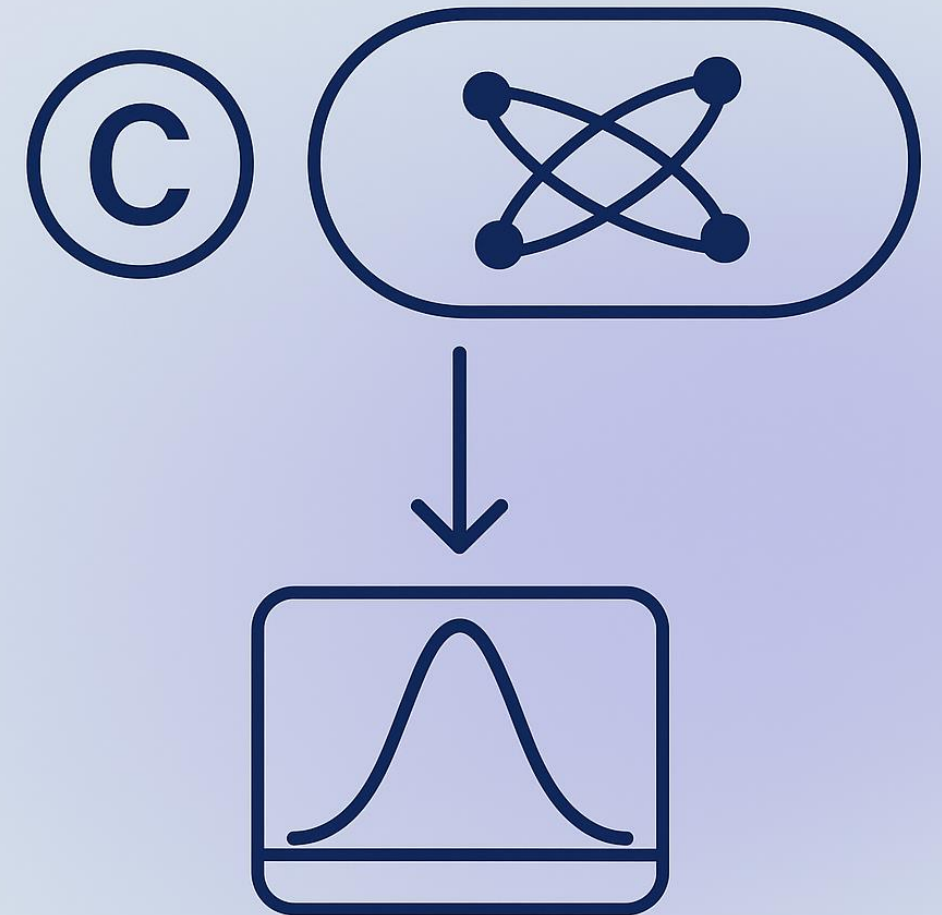
# **SVM: мощное разделение классов**

Машина опорных векторов (SVM) ищет оптимальную гиперплоскость для разделения классов.

Метод способен эффективно работать даже с нелинейно разделимыми данными.

# Настройка SVM в RapidMiner

- C (штраф за ошибки классификации)
  - Тип ядра (линейное или RBF)
  - Gamma (параметр ядра)
- Требует нормализации признаков для корректной работы.





# Важность нормализации данных

## NORMALIZATION

Standardization ▼

Standardization

Define range

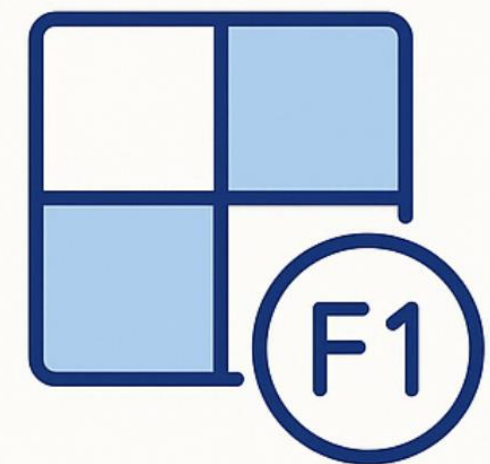
Interquartile range

Нормализация приводит все признаки к единому масштабу (0-1 или среднее=0).

Особенно важна для k-NN и SVM, где масштаб признаков сильно влияет на результаты.

# Метрики качества классификации

- Accuracy (общая точность классификации)
- Precision (точность позитивных предсказаний)
- Recall (доля обнаруженных позитивных примеров)
- F1-score (баланс между Precision и Recall)







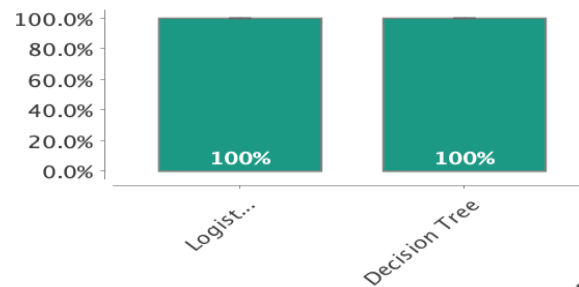
# Сравнение моделей в RapidMiner

## Results

- Comparison
  - Overview
  - ROC Comparison
- Logistic Regression
  - Model
  - Weights
  - Simulator
  - Performance
  - Lift Chart
  - Predictions
  - Production Model
- Decision Tree
- Random Forest

## Overview

### Accuracy



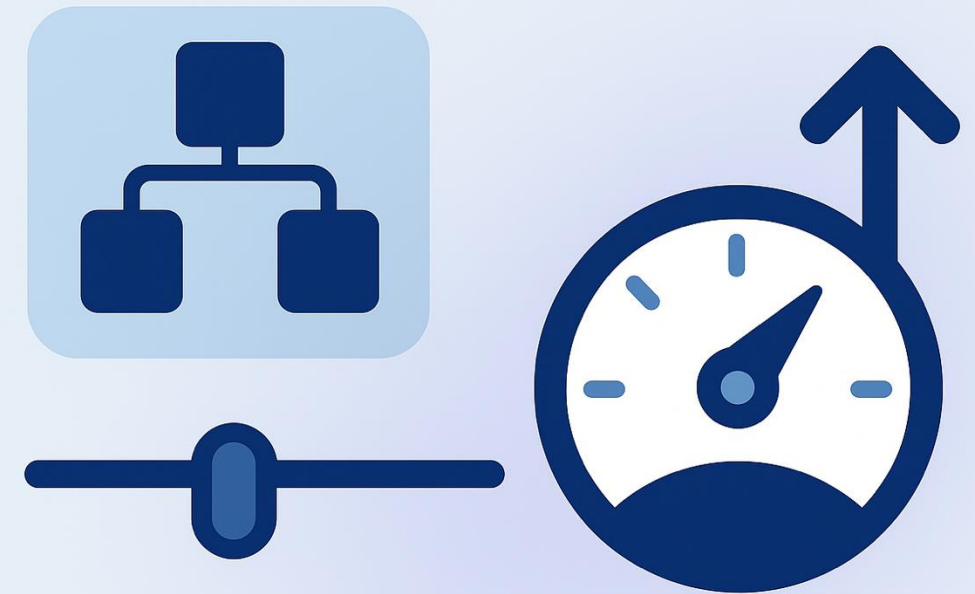
Accuracy	Accuracy
Accuracy	100.0%
Classification Error	100.0%
AUC	100.0%
Precision	100.0%
Recall	100.0%
F Measure	100.0%
Sensitivity	98.2%
Specificity	

Используя Cross Validation, оценивают каждую модель по метрикам Accuracy, Precision, Recall и F1-score.

Это позволяет выбрать оптимальный алгоритм для конкретной задачи.

# Влияние параметров моделей на качество

Правильный подбор гиперпараметров (число соседей  $k$ , глубина деревьев, ядро SVM) может значительно повысить точность и устойчивость модели.



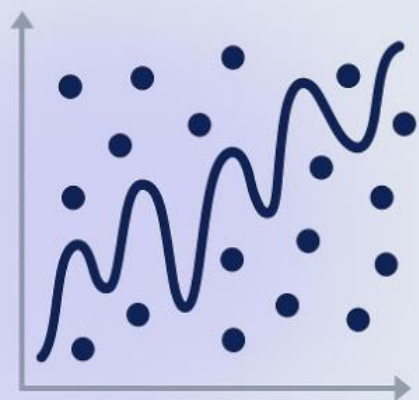


# Переобучение и его признаки

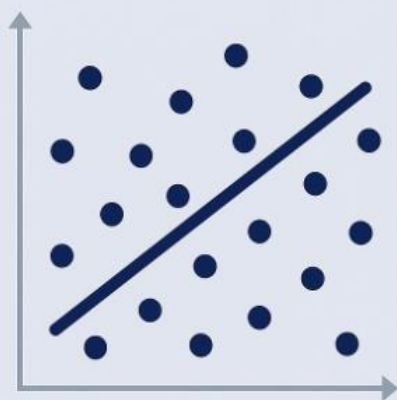
Переобучение возникает, когда модель слишком сильно адаптируется к обучающим данным и плохо обобщает новые примеры.

Симптом: высокая точность на обучении, низкая на тестировании.

overfitting

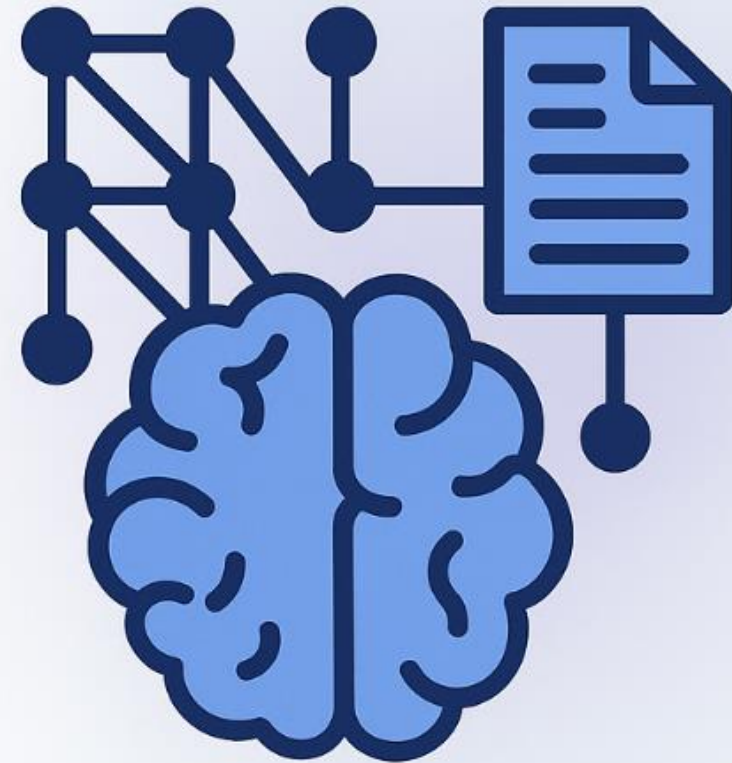


underfitting



# Способы борьбы с переобучением

- Упрощение модели (уменьшение глубины дерева)
- Регуляризация (ограничение параметров)
- Использование Cross Validation для объективной оценки
- Увеличение объема обучающих данных





## Заключение и рекомендации

Использование нескольких ML-моделей и Cross Validation в RapidMiner позволяет объективно выбрать наиболее подходящий метод, оптимизировать качество прогнозов и избегать распространённых ошибок в машинном обучении.