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

Кафедра информационной безопасности киберфизических систем

Москва 2025

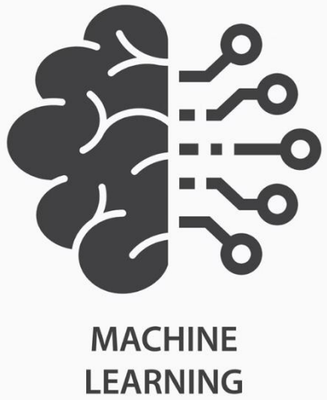
Изучение возможностей ML-моделей

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?

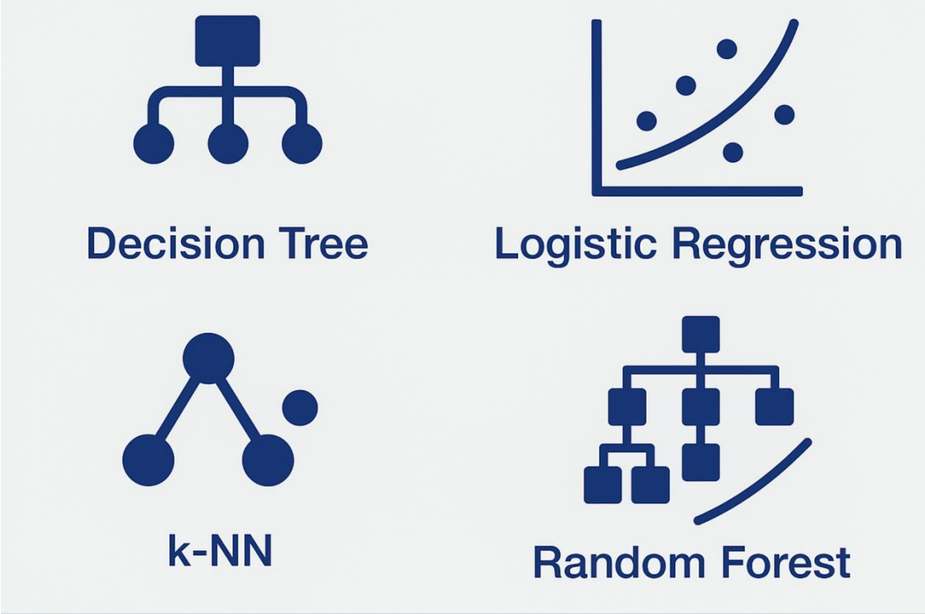


Использование нескольких моделей важно для:

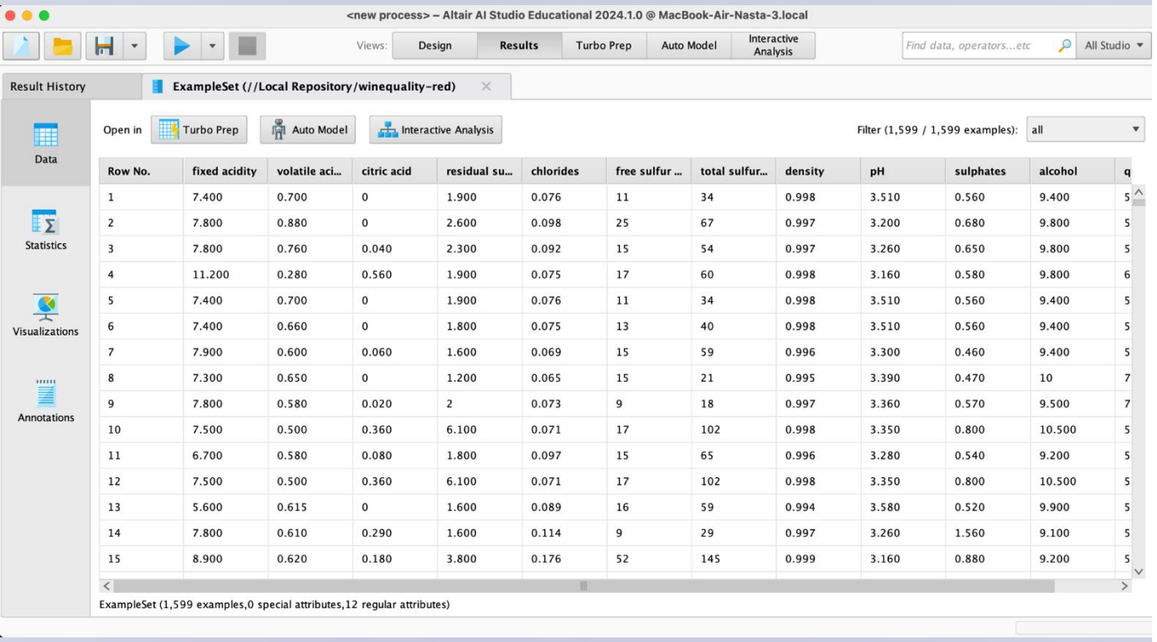
- Подбора лучшего подхода к конкретной задаче
- Учёта разных типов данных
- Сравнения эффективности и точности различных методов обучения

Обзор используемых алгоритмов

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



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».

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.

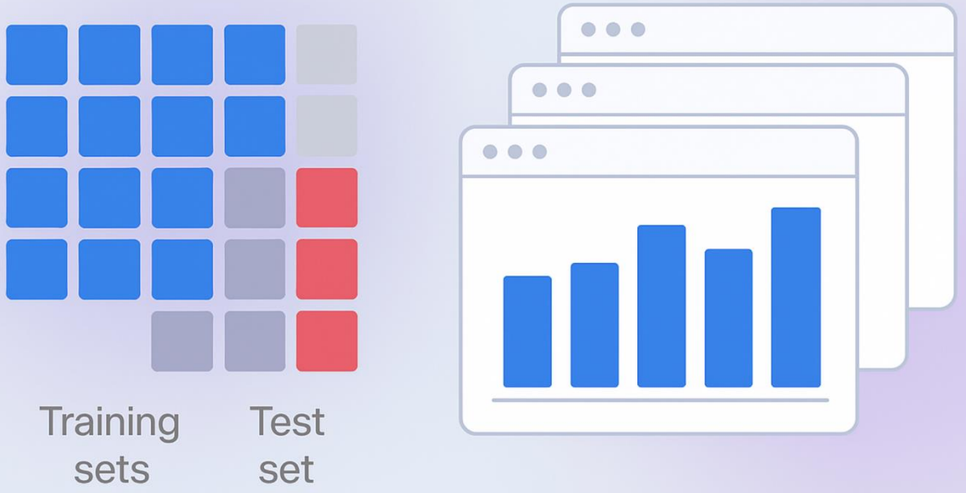
Основные этапы анализа в RapidMiner

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

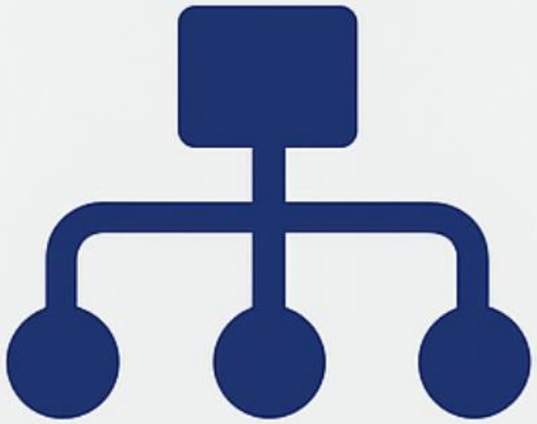
Что такое Cross Validation?

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

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



Decision Tree: принцип работы



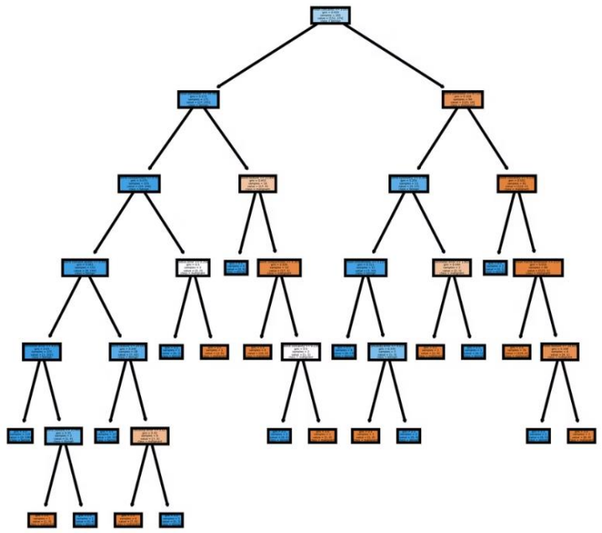
Модель разбивает данные на группы вопросами («pH > 3.2?», «alcohol > 10?»).

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

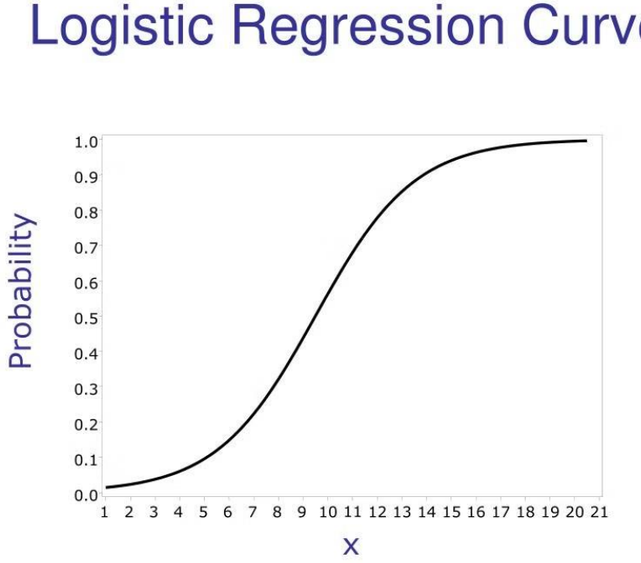
Логистическая регрессия оценивает вероятность принадлежности объекта к классу, используя линейную комбинацию признаков.

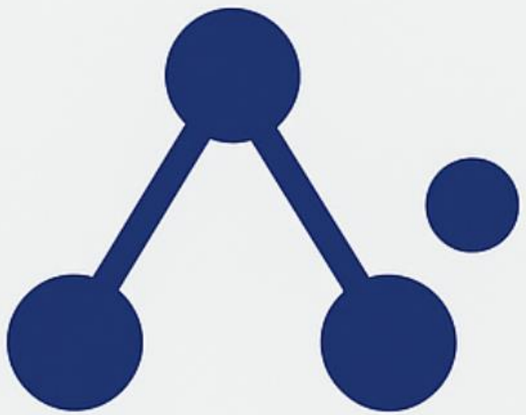
Это простой и эффективный подход, часто используемый в базовых задачах классификации.

Logistic Regression: интерпретация

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

Положительные коэффициенты увеличивают вероятность, отрицательные — уменьшают.





k-NN: принцип классификации

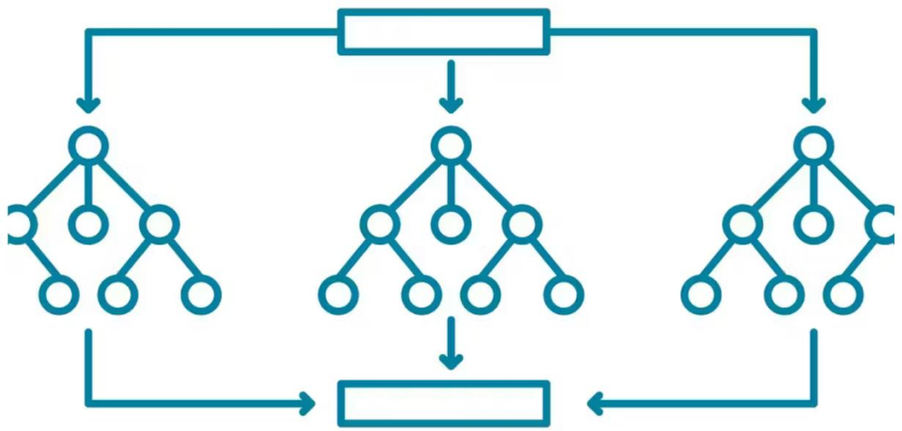
Метод k ближайших соседей относит объект к тому классу, который преобладает среди ближайших k объектов.

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



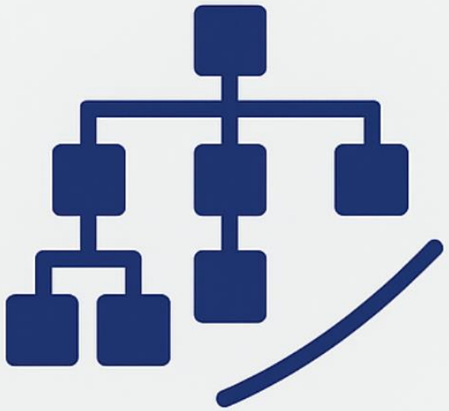
Random Forest обучает множество деревьев решений на случайных подвыборках данных и признаков, затем агрегирует результаты.

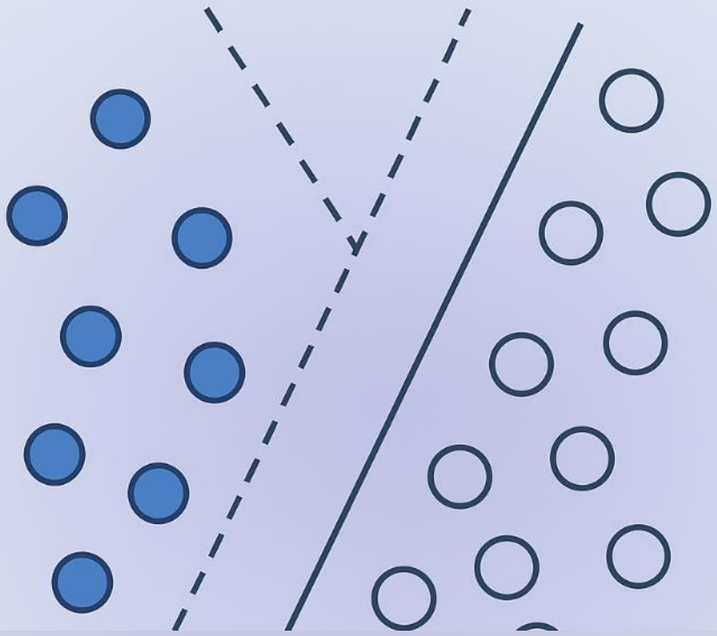
Это уменьшает риск переобучения и повышает стабильность модели.

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.

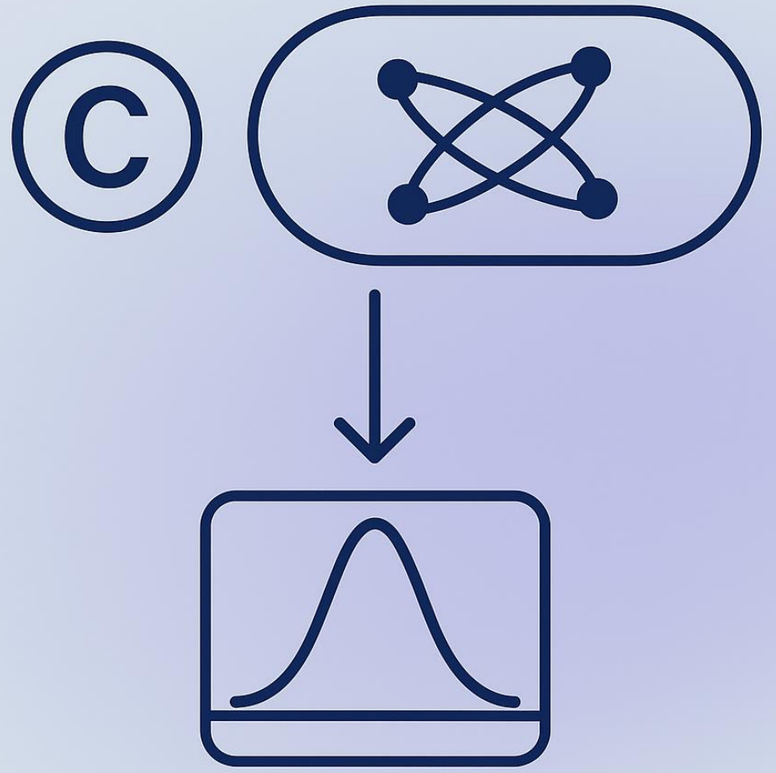
Настройка SVM в RapidMiner

- C (штраф за ошибки классификации)

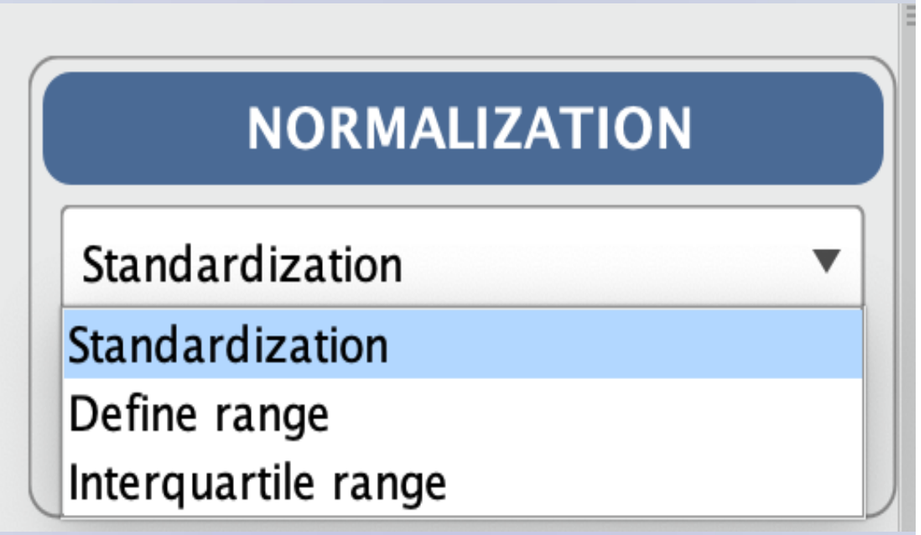
- Тип ядра (линейное или RBF)

- Gamma (параметр ядра)

Требует нормализации признаков для корректной работы.



Важность нормализации данных

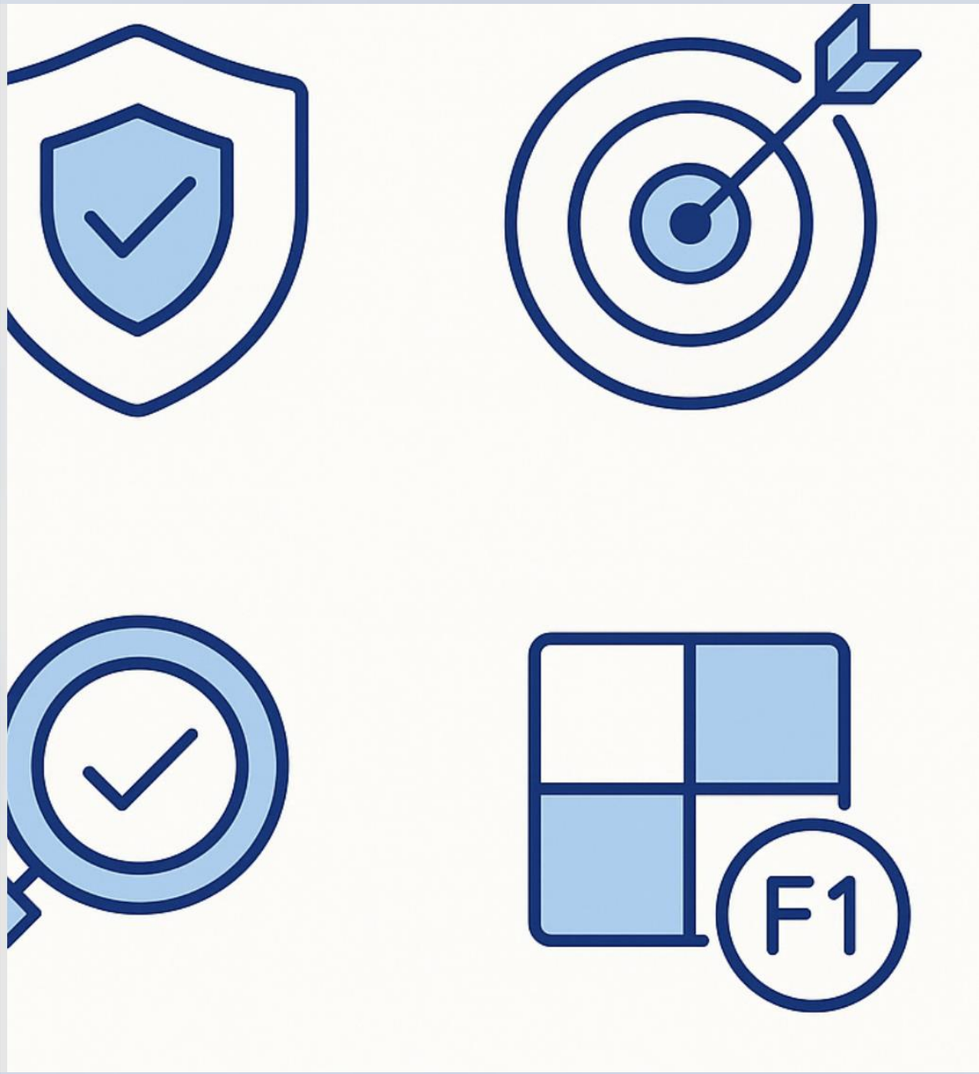


Нормализация приводит все признаки к единому масштабу (0-1 или среднее=0).

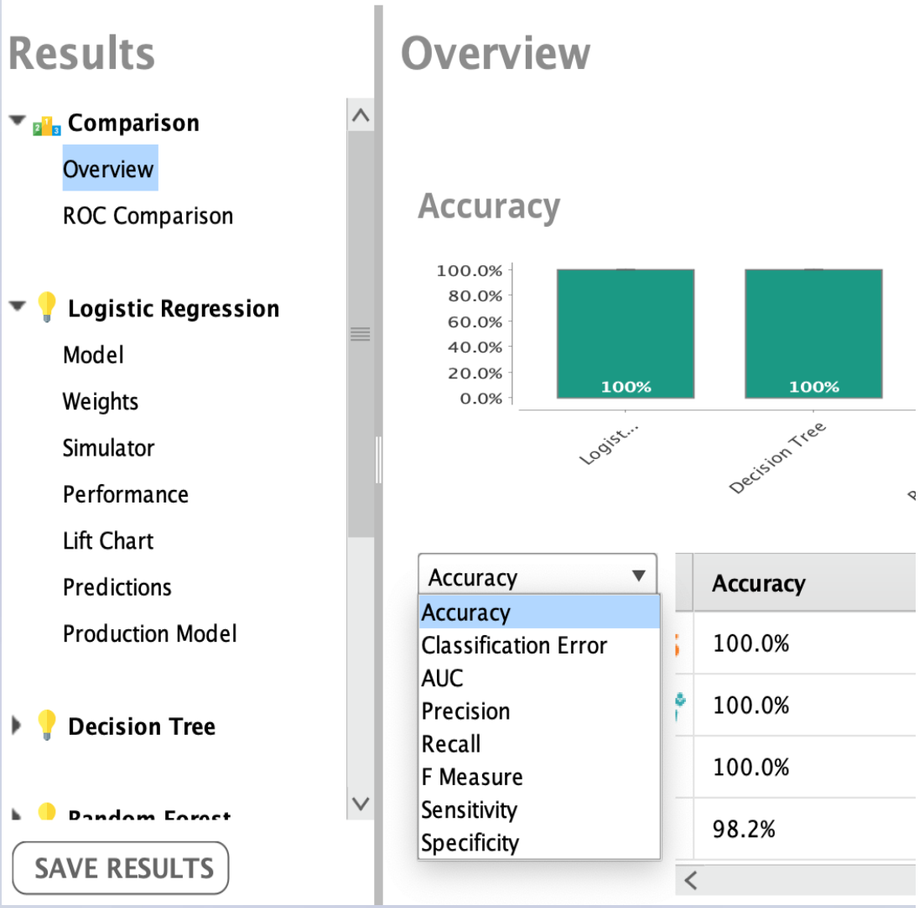
Особенно важна для k-NN и SVM, где масштаб признаков сильно влияет на результаты.

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

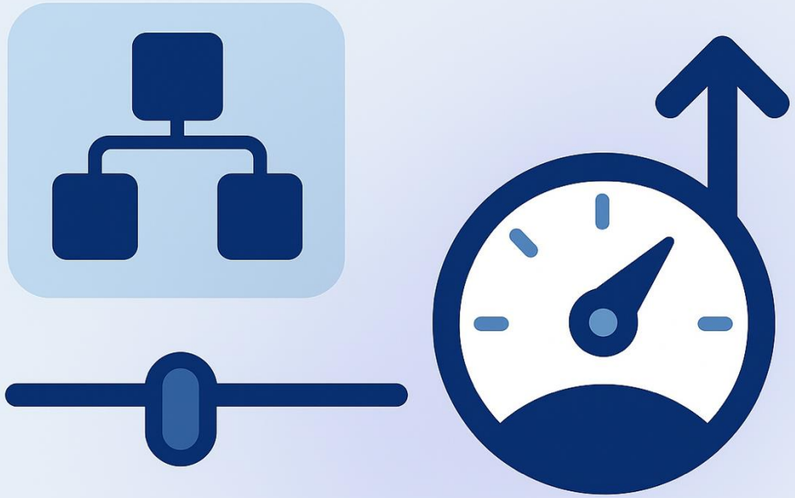


Используя Cross Validation, оценивают каждую модель по метрикам Accuracy, Precision, Recall и F1-score.

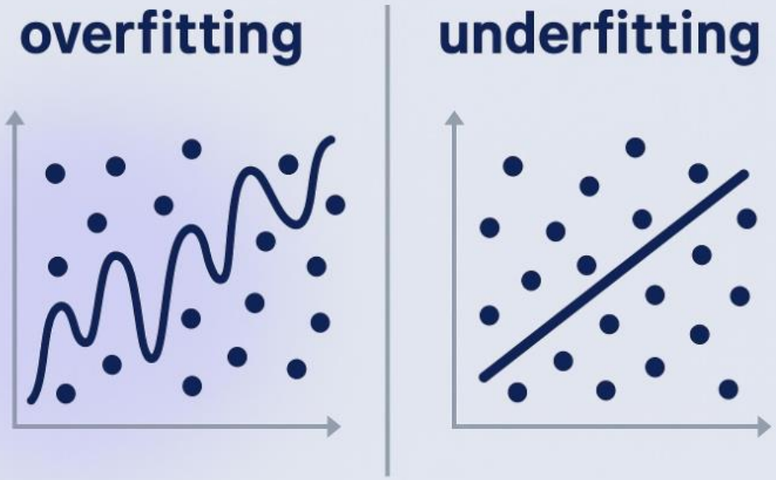
Это позволяет выбрать оптимальный алгоритм для конкретной задачи.

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

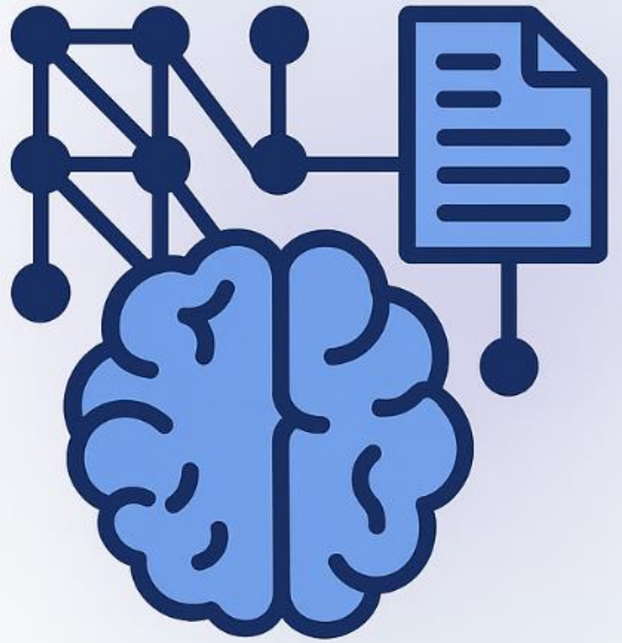


Переобучение возникает, когда модель слишком сильно адаптируется к обучающим данным и плохо обобщает новые примеры.

Симптом: высокая точность на обучении, низкая на тестировании.

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.