Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
(НИУ ВШЭ)

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

ПРАКТИЧЕСКАЯ РАБОТА № 6

ТЕМА РАБОТЫ
«Изучение возможностей ML-моделей в RapidMiner»

Москва, 2025

# Практическая работа 6: Изучение возможностей ML-моделей в RapidMiner



## Цель работы

Познакомиться с возможностями построения моделей машинного обучения в RapidMiner с использованием стандартных операторов (без AutoModel). Студенты научатся:

- Выбирать различные алгоритмы машинного обучения (Decision Tree, Logistic Regression, k-NN, Random Forest, SVM).
- Настраивать параметры моделей.
- Применять операторы для обучения, валидации и оценки качества моделей.
- Сравнивать результаты разных моделей и интерпретировать полученные метрики.

## Целевая аудитория

Студенты, уже имеющие базовое представление о RapidMiner, умеющие работать с данными, выполнять предобработку и визуализацию. После выполнения студенты будут готовы самостоятельно применять операторы для обучения моделей, выбирать параметры и интерпретировать результаты.

## Идея и концепция

Для демонстрации возможностей различных моделей будет использован набор данных **"Wine Quality"**. Набор данных содержит химические характеристики красных вин (кислотность, сахар, pH, содержание алкоголя и др.) и оценку качества (quality) по шкале от 0 до 10. Задача – предсказать качество вина (как категориальную или упрощенную бинарную метку: высокое/низкое качество) на основе его химического профиля.

Используя этот датасет, студенты смогут:

- Взять признаки вина (числовые данные) и построить несколько моделей классификации.
- Настроить параметры моделей (например, глубину дерева для Decision Tree, количество соседей для k-NN, гиперпараметры SVM).
- Применить кросс-валидацию для более надежной оценки качества моделей.
- Сравнить точность, полноту, F1-score разных алгоритмов и сделать вывод, какой из них лучше подходит для данной задачи.

## Содержание практической работы

### Введение в RapidMiner

Краткое напоминание о работе с процессами в RapidMiner: использование операторов, подключение операторов друг к другу, просмотр результатов. Акцент на том, что в данной работе не используется AutoModel – все действия по построению и оценке моделей будут выполнены вручную, с помощью операторов из библиотек RapidMiner.

- Установите RapidMiner Studio (если не установлено).
- Ознакомьтесь с панелью операторов, найдите разделы с моделями (Modeling → Predictive → …).

- Проверьте, что у вас есть датасет “winequality-red.csv”.

### О наборе данных и задаче работы

Датасет о качестве красного вина (winequality-red.csv) содержит столбцы:

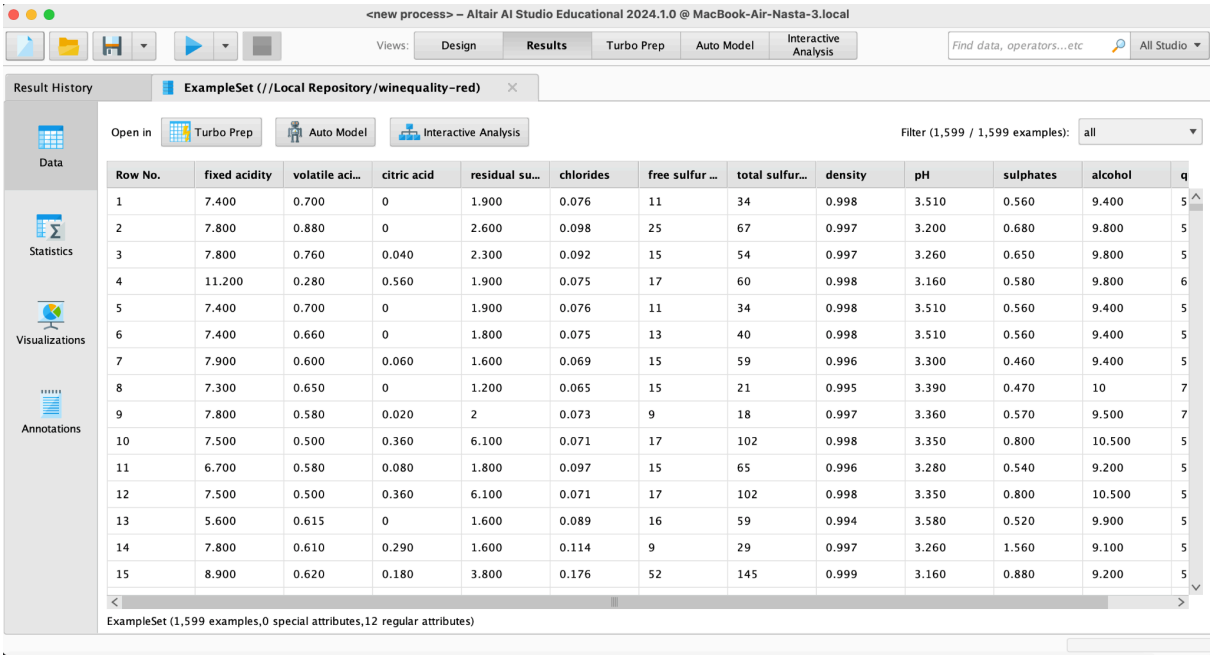
- Химические параметры вина: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
- Целевой признак: quality (оценка качества от 0 до 10).

Цель: предсказать качество вина, упростив задачу до классификации – например, вина с оценкой 6 и выше считать «хорошими» (class = good), а ниже 6 – «не очень хорошими» (class = not_good). Такое преобразование будет сделано на этапе предобработки данных.

### Работа с данными

#### Загрузка набора данных

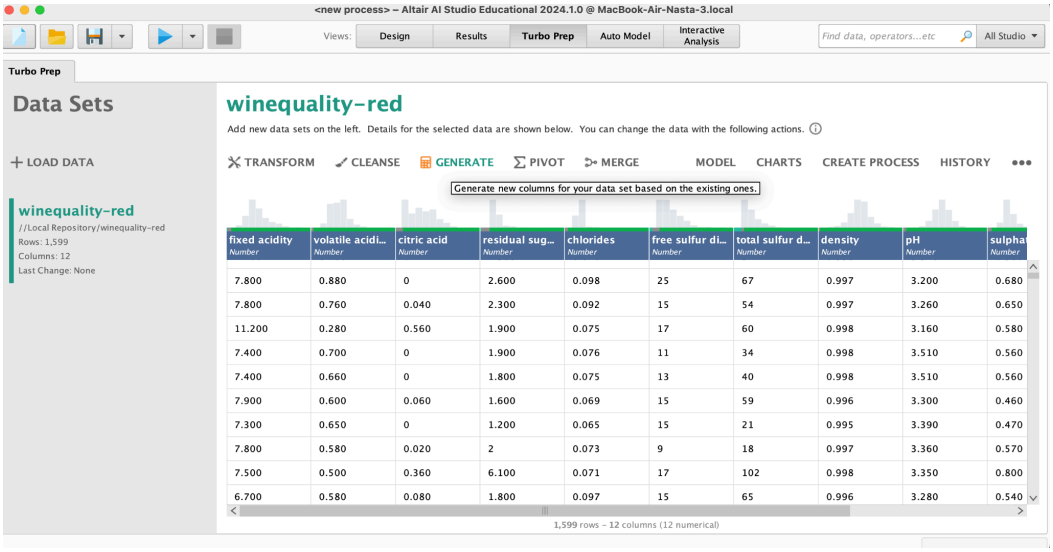
- Откройте RapidMiner Studio.
- Выберите "Create New Process".
- Перетащите оператор **"Read CSV"** в рабочее пространство.
- Загрузите файл "winequality-red.csv" из вашего локального хранилища.
- Подключите выход “Read CSV” к “Result” и нажмите “Run”, чтобы просмотреть данные.




| 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 |

*рис. 1: Просмотр датасета после загрузки*

## Предобработка данных (TurboPrep или ручная настройка)




| fixed acidity Number | volatile acidi... Number | citric acid Number | residual sug... Number | chlorides Number | free sulfur di... Number | total sulfur d... Number | density Number | pH Number | sulpha Number |
|---|---|---|---|---|---|---|---|---|---|
| 7.800 | 0.880 | 0 | 2.600 | 0.098 | 25 | 67 | 0.997 | 3.200 | 0.680 |
| 7.800 | 0.760 | 0.040 | 2.300 | 0.092 | 15 | 54 | 0.997 | 3.260 | 0.650 |
| 11.200 | 0.280 | 0.560 | 1.900 | 0.075 | 17 | 60 | 0.998 | 3.160 | 0.580 |
| 7.400 | 0.700 | 0 | 1.900 | 0.076 | 11 | 34 | 0.998 | 3.510 | 0.560 |
| 7.400 | 0.660 | 0 | 1.800 | 0.075 | 13 | 40 | 0.998 | 3.510 | 0.560 |
| 7.900 | 0.600 | 0.060 | 1.600 | 0.069 | 15 | 59 | 0.996 | 3.300 | 0.460 |
| 7.300 | 0.650 | 0 | 1.200 | 0.065 | 15 | 21 | 0.995 | 3.390 | 0.470 |
| 7.800 | 0.580 | 0.020 | 2 | 0.073 | 9 | 18 | 0.997 | 3.360 | 0.570 |
| 7.500 | 0.500 | 0.360 | 6.100 | 0.071 | 17 | 102 | 0.998 | 3.350 | 0.800 |
| 6.700 | 0.580 | 0.080 | 1.800 | 0.097 | 15 | 65 | 0.996 | 3.280 | 0.540 |

*рис. 2: Открытие датасета во вкладке "TurboPrep«*

- Перейдите в TurboPrep или используйте оператор “Generate”.
- Создайте бинарный целевой признак:
- Используя оператор "Generate", создайте новый столбец, например “quality_class”, по правилу:
    - quality_class = if(quality >= 6, "good", "not_good")
- Удалите исходный столбец “quality” или оставьте для справки. Главное – целевой признак теперь категориальный.

- Примените “Set Role” к столбцу “quality_class”, назначив его как “label”.

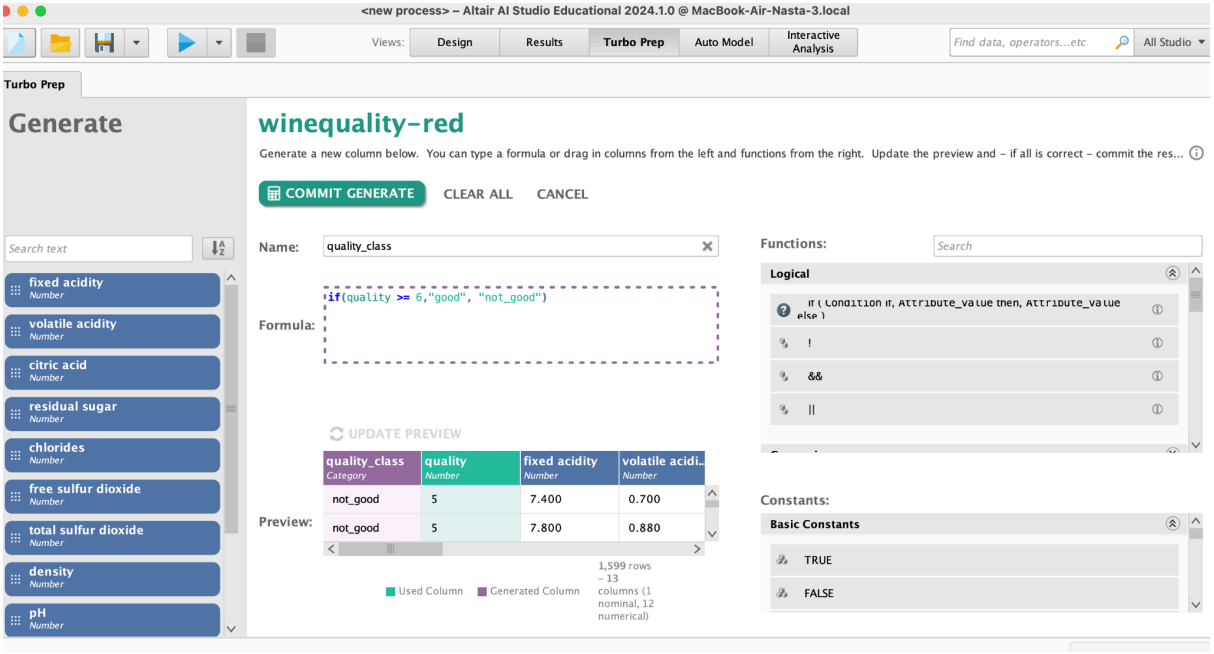



*рис. 3: Использование оператора "Generate" для создания нового столбца*

## Подготовка данных для моделирования

- При необходимости нормализуйте признаки (оператор "Normalize") для улучшения работы алгоритмов k-NN или SVM.
- Убедитесь, что все нужные признаки числовые, а целевой – категориальный.

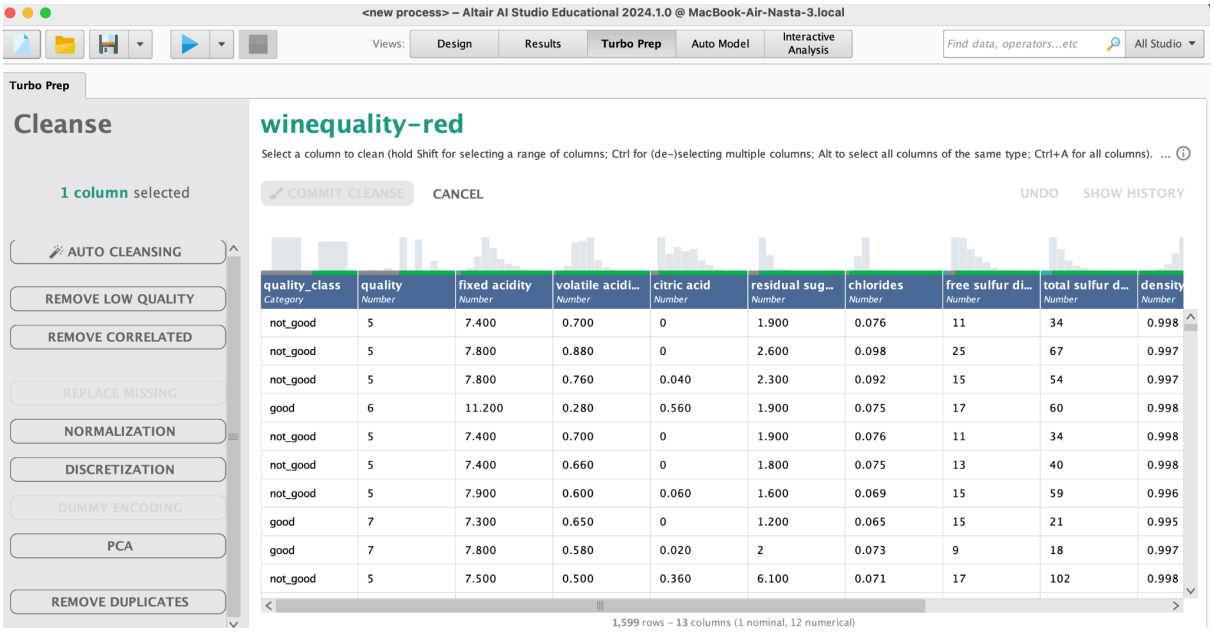



*рис. 4: Использование оператора "Normalization" для масштабирования столбцов*

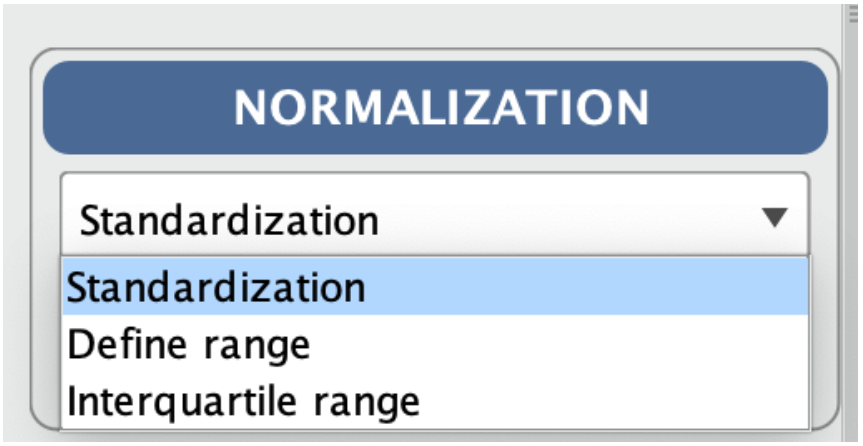



*рис. 5: Выбор способа масштабирования данных внутри оператора "Normalization"*

## Построение ML-моделей

В этой работе не используем AutoModel. Вместо этого применяем операторы из библиотеки RapidMiner.

1. **Decision Tree**

- Добавьте оператор **"Decision Tree"**.

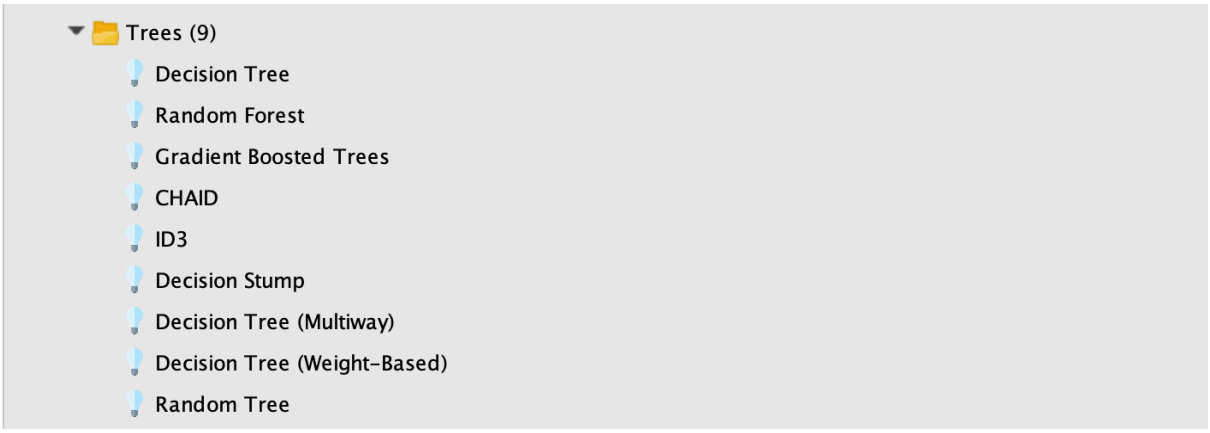



*рис. 6: Выбор оператора "Decision Tree"*

- Подключите к нему выход данных (ExampleSet) с обработанными данными.
- Подключите выход Decision Tree к оператору (для последующей оценки). Но сначала нам нужен механизм проверки качества – используем кросс-валидацию.

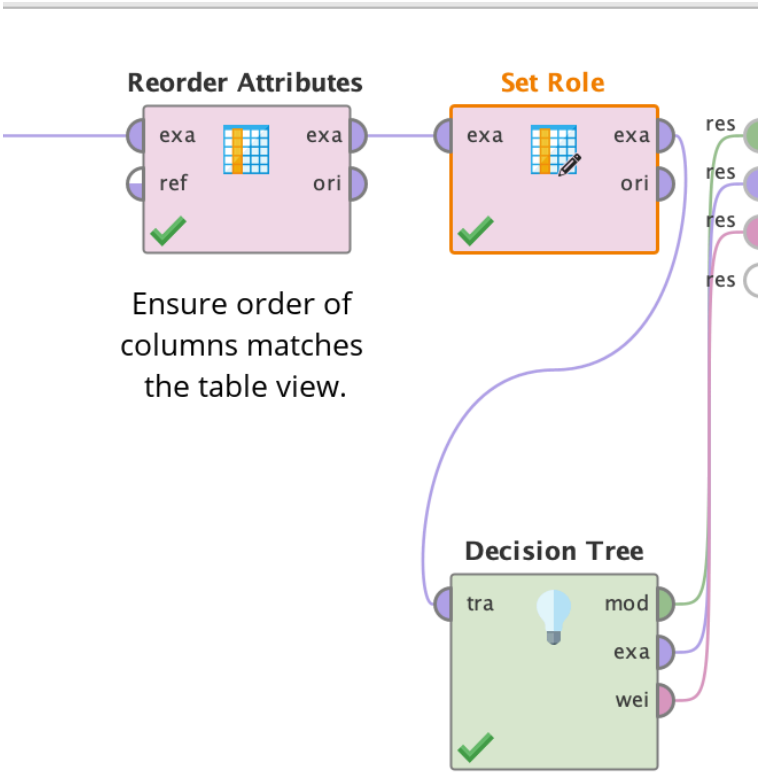



*рис. 7: Включение оператора "Decision Tree" в основную цепочку*

## 2. Кросс-валидация (Cross Validation)

- Добавьте оператор **"Cross Validation"** на рабочее пространство.
- Внутри оператора Cross Validation поместите:
    - В “Training” часть – оператор “Decision Tree” (и его вход – данные).
    - В “Testing” часть – оператор “Apply Model” и затем оператор “Performance (Classification)”.

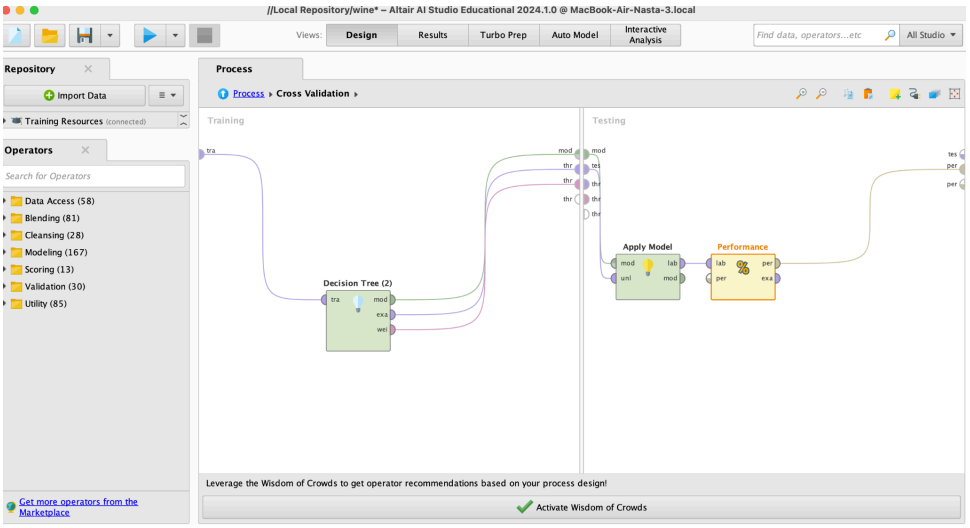



*рис. 8: Настройка кросс-валидациия*

- Cross Validation будет принимать данные извне и делить их на фолды.
- Примените те же шаги для других моделей, заменяя Decision Tree на другие алгоритмы (ниже).

Decision Tree – Weights

| Attribute | Weight |
|---|---|
| quality | 0.353 |
| alcohol | 0.077 |
| density | 0.063 |
| citric acid | 0.048 |
| fixed acidity | 0.043 |
| total sulfur dioxide | 0.036 |
| residual sugar | 0.026 |
| free sulfur dioxide | 0.016 |
| volatile acidity | 0.009 |
| chlorides | 0.008 |
| sulphates | 0.007 |
| pH | 0.006 |

*рис. 9: Просмотр весов переменных, рассчитанных после обучения модели "Дерево решений"*

## 3. Logistic Regression

- Аналогично, используйте оператор **"Logistic Regression"** вместо Decision Tree в части Training.

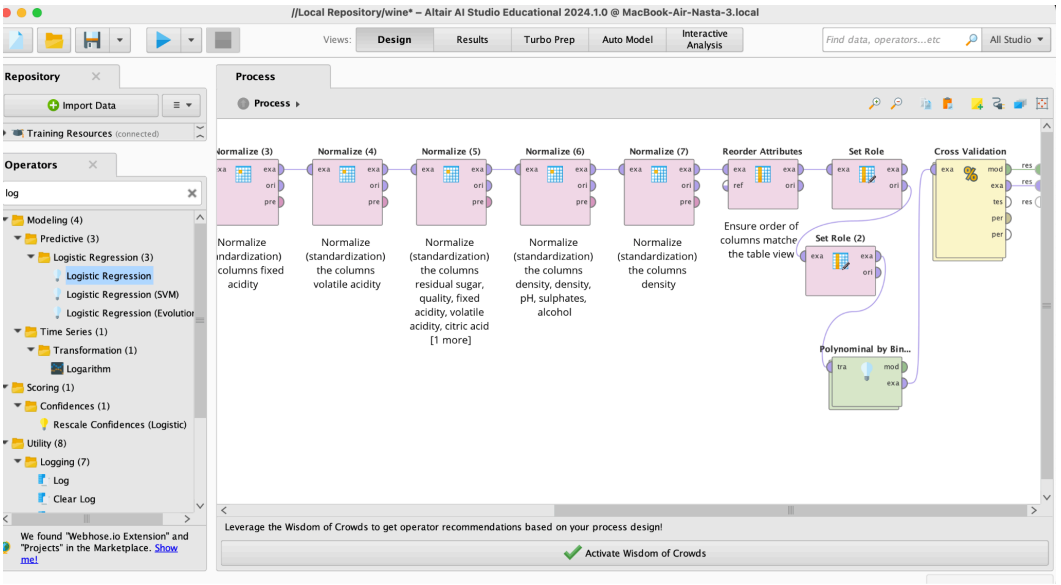



*рис. 10: Выбор Logistic Regression на панели слева*

- Оцените точность модели с помощью кросс-валидации.

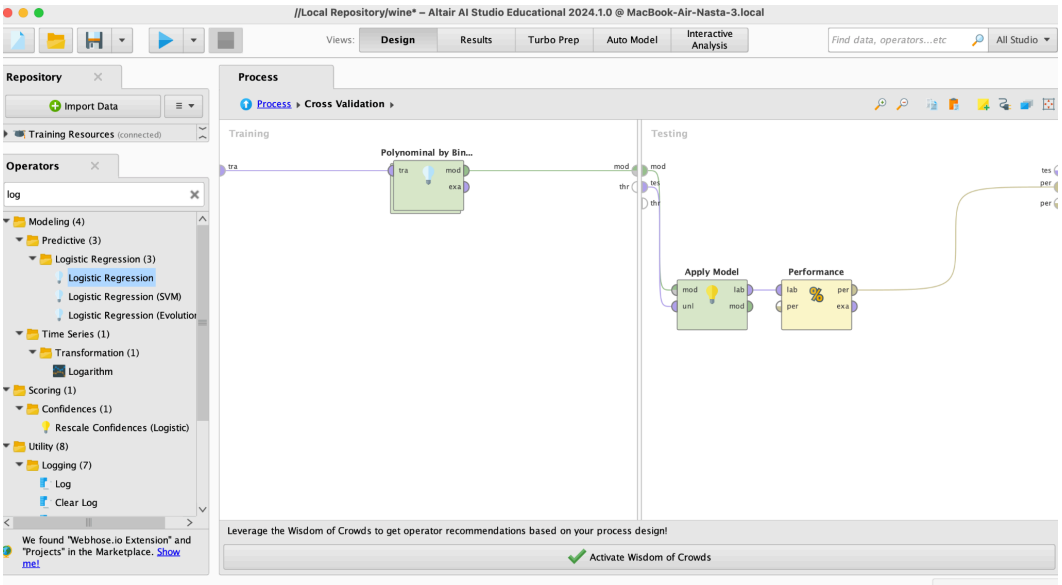



*рис. 11: Настройка кросс-валидации для модели логистической регрессии*

Logistic Regression – Model

| Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value |
|---|---|---|---|---|---|
| quality | 26.246 | 26.299 | 42.944 | 0.611 | 0.541 |
| fixed acidity | -0.063 | -0.064 | 95.969 | -0.001 | 0.999 |
| volatile acidity | -0.024 | -0.024 | 47.636 | -0.000 | 1.000 |
| citric acid | -0.046 | -0.046 | 62.494 | -0.001 | 0.999 |
| residual sugar | -0.070 | -0.056 | 54.825 | -0.001 | 0.999 |
| chlorides | 1.635 | 0.071 | 946.188 | 0.002 | 0.999 |
| free sulfur dioxide | 0.005 | 0.058 | 4.631 | 0.001 | 0.999 |
| total sulfur dioxide | -0.006 | -0.191 | 1.657 | -0.004 | 0.997 |
| density | 0.096 | 0.093 | 87.394 | 0.001 | 0.999 |
| pH | 0.005 | 0.005 | 64.011 | 0.000 | 1.000 |
| sulphates | 0.006 | 0.006 | 41.394 | 0.000 | 1.000 |

*рис. 12: Просмотр рассчитанных статистических показателей данных после применения логистической регрессии*

4. **k-Nearest Neighbors (k-NN)**

- Добавьте оператор **"k-NN"**.

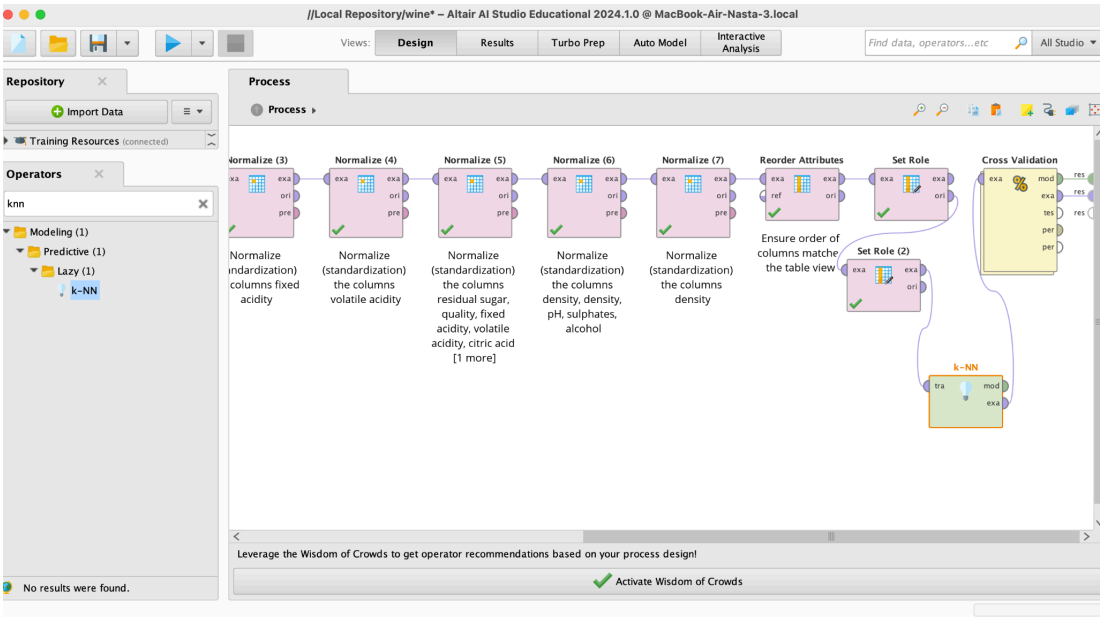



*рис. 13: Выбор оператора k-NN на левой панели*

- Настройте параметр k (например, k=5 или k=7) и оцените качество через кросс-валидацию.

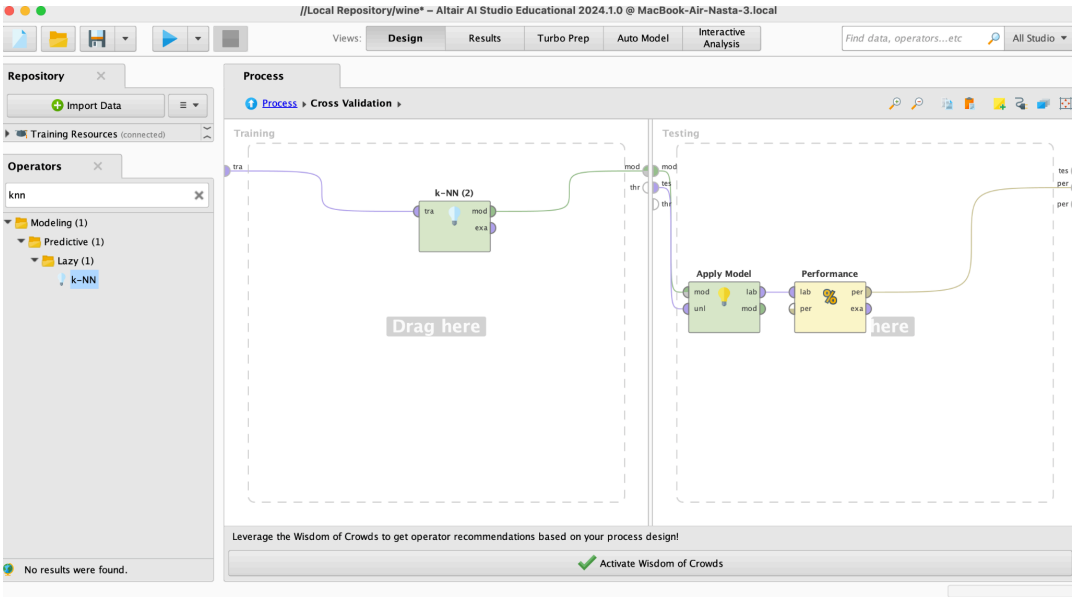



*рис. 14: Настройка кросс-валидации для k-NN*

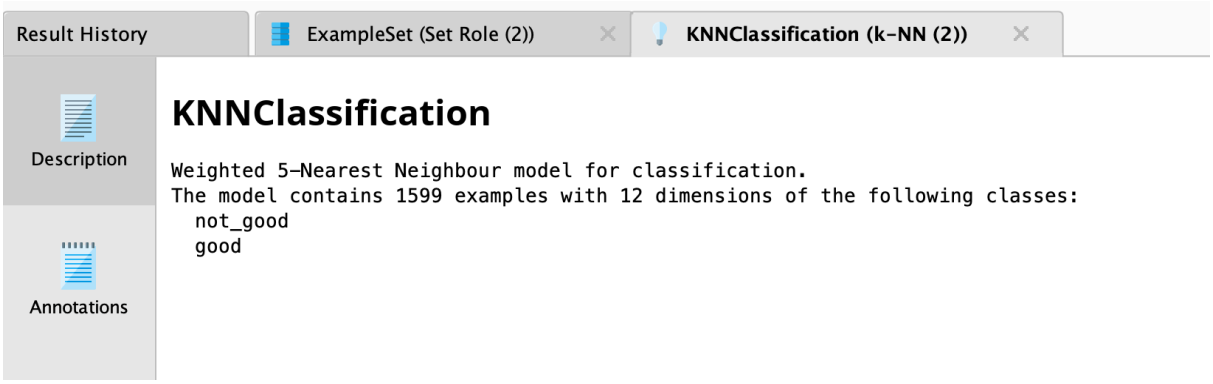



*рис. 15: Просмотр модели KNNClassification*

5. **Random Forest**

- Оператор **"Random Forest"** позволит построить ансамблевую модель.
- Настройте количество деревьев (number of trees), глубину (maximal depth). Снова оцените точность модели с помощью кросс-валидации.

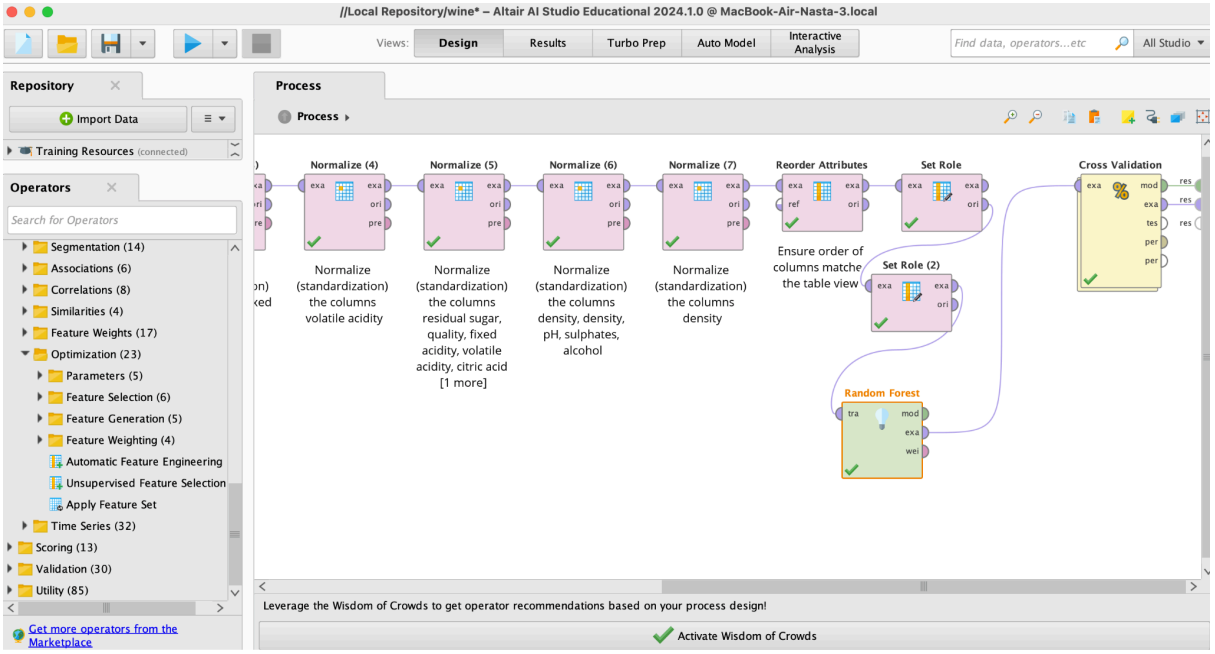



*рис. 16: Включение оператора Random Forest в цепочку операторов*

6. **Support Vector Machine (SVM)**

- Используйте оператор **"SVM"**.
- Оцените качество модели.

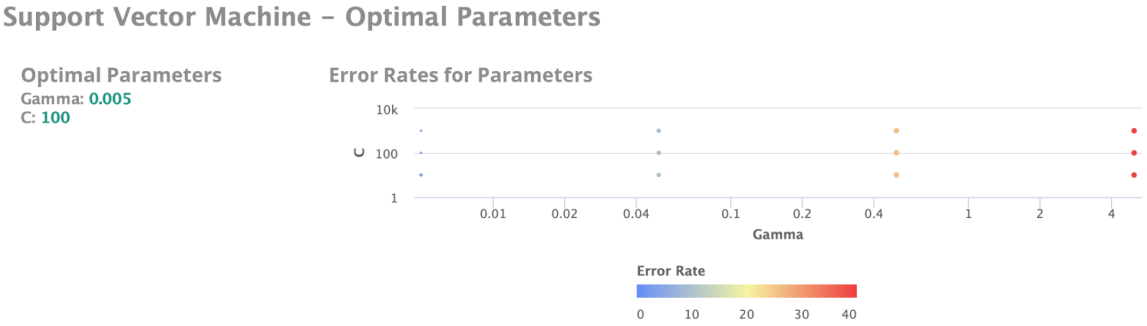



*рис. 17: Оптимальные параметры с помощью модели SVM*

Support Vector Machine – Model

**Kernel Model**

```
Total number of Support Vectors: 228
Bias (offset): -0.059

w[quality] = -128.590
w[fixed acidity] = 33.650
w[volatile acidity] = 98.008
w[citric acid] = -19.373
w[residual sugar] = -99.441
w[chlorides] = 73.143
w[free sulfur dioxide] = 15825.282
w[total sulfur dioxide] = 47858.892
w[density] = 91.172
w[pH] = -87.450
w[sulphates] = -173.511
w[alcohol] = -263.440

number of classes: 2
number of support vectors for class not_good: 130
number of support vectors for class good: 98
```

-

*рис. 18: Просмотр рассчитанных статистических показателей данных после применения SVM-модели*

## Сравнение результатов

- Для каждой модели выполните кросс-валидацию и зафиксируйте метрики: Accuracy, Precision, Recall, F1-score.
- Сравните результаты в выходном окне “Performance” или сохраните значения в таблицу.

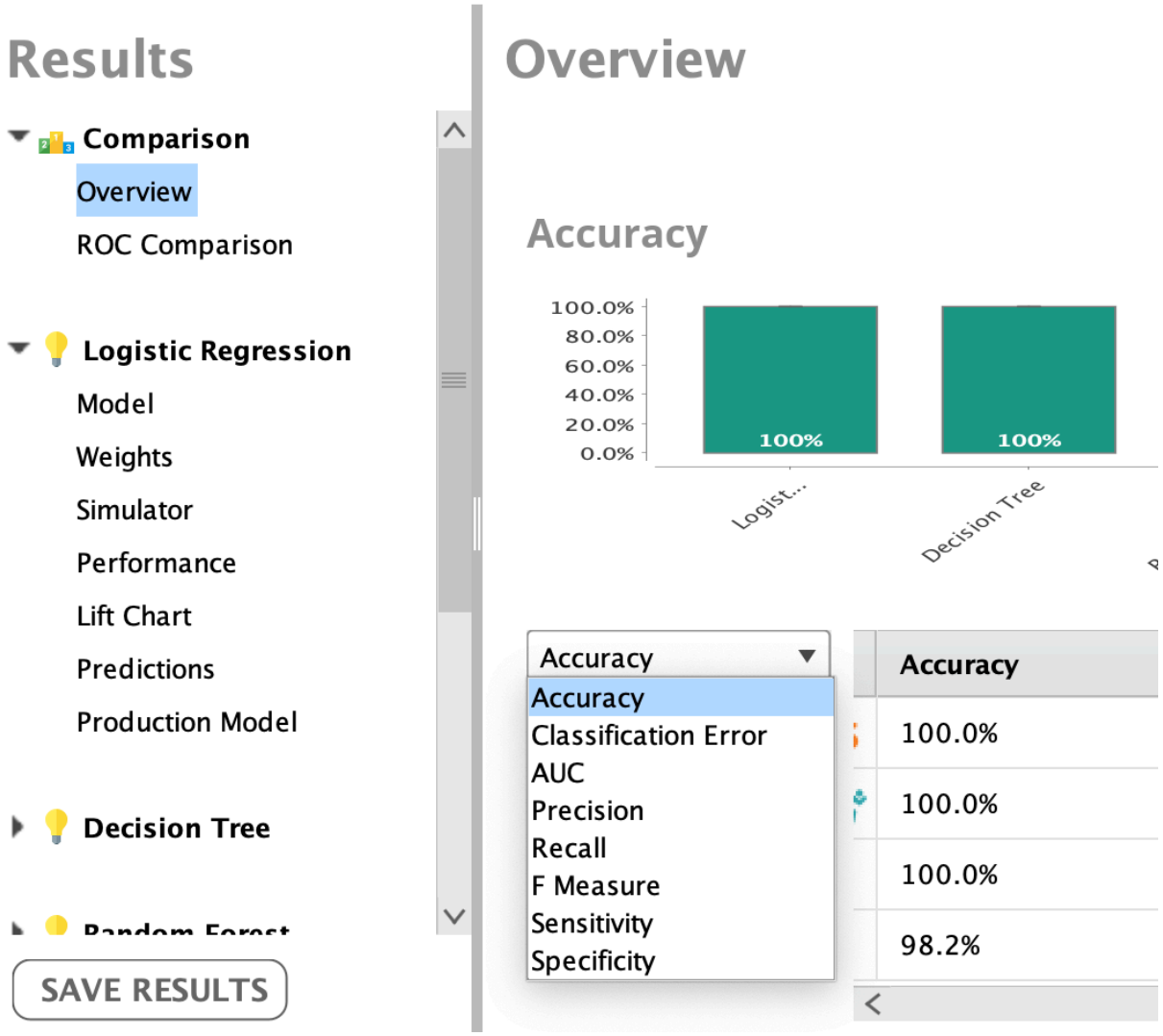



*рис. 19: Сравнения результатов всех использованных моделей*

- Сделайте вывод, какая модель работает лучше для предсказания качества вина.

### Анализ результатов

- Проверьте, сильно ли различаются результаты разных моделей.
- Посмотрите на метрику F1-score, если классы несбалансированные.
- Подумайте, влияет ли нормализация признаков на результаты k-NN или SVM.

## Приобретенные навыки

- Умение самостоятельно обучать модели в RapidMiner без AutoModel, используя стандартные операторы.

- Навык настройки параметров алгоритмов (Decision Tree, k-NN, Random Forest, SVM).
- Понимание процесса кросс-валидации и её необходимости.
- Способность интерпретировать метрики качества (accuracy, precision, recall, F1-score) и использовать их для сравнения моделей.
- Опыт принятия решений о выборе модели на основе метрик производительности.

## Обобщенная задача для выполнения индивидуального варианта

Предложенный в индивидуальном варианте набор данных содержит числовые признаки и категориальную целевую переменной. В ходе данной работы вам предлагается построить в RapidMiner (Altair AI Studio) несколько моделей машинного обучения и провести их сравнительный анализ:

Загрузка и подготовка данных:

- Импортируйте данные через оператор Read CSV.
- Обработайте пропуски (Replace Missing Values), при необходимости преобразуйте типы признаков (Nominal–Numerical).
- Создайте целевой признак: если исходная метка числовая, сведите ее к категориям (Generate Attributes или TurboPrep).
- Разделите выборку на train/test или используйте оператор Cross Validation.

Обучение моделей:

- Постройте не менее трех моделей классификации (минимум один алгоритм из каждой группы): дерево решений (Decision Tree), линейная модель (Logistic Regression), ансамблевая модель (Random Forest), метод опорных векторов (SVM), k-ближайших соседей (k-NN).
- Для каждой модели настройте ключевые гиперпараметры (глубину дерева, число соседей, количество деревьев в ансамбле и т. д.).

Оценка и сравнение:

- Используйте Cross Validation или разделение train/test для оценки качества.

- Соберите метрики: Accuracy, Precision, Recall, F1-score.
- Сформируйте сводную таблицу с результатами всех моделей.

Выводы и рекомендации:

- Проанализируйте, какая модель показала наилучшие метрики и почему.
- Оцените влияние настройки гиперпараметров на качество.

## Распределение вариантов