**Кейс «Анализ факторов, влияющих на категорию полученного осадка на химическом заводе**

**Аннотация.** На крупном химическом заводе для получения целевого продукта в большом объеме используют химические реакторы. Перед запуском производства в промышленных масштабах в химической лаборатории проводят реакцию двух органических веществ. В зависимости от множества условий, в ходе реакции может образоваться однородный осадок (полезный продукт), полностью состоящий из изомера №1, либо из изомера №2. Также необходимаяреакция может просто не произойти. В таком случае осадок будет состоять из смеси неопределенных веществ, непригодной для дальнейшей работы (брак). Определить, что получилось в реакторе можно лишь проведя химический анализ. Причем для определения каждого изомера полезного продукта требуется отдельный дорогостоящий индикатор. Если лаборант с первого раза подтверждает наличие одного из двух изомеров, дальнейшие тесты не требуются. Но в большинстве случаев с выбором индикатора ошибаются и приходится проводить еще один анализ.

В связи с этим, возникла необходимость создания модели машинного обучения, которая подсказывала бы лаборантам, какой индикатор стоит использовать в первую очередь. Достаточная точность модели позволит уменьшить вероятность того, что потребуется проводить дополнительный анализ полученного вещества.

**Цель:** построить классификатор, который предсказывает категорию полученного осадка (изомер №1, изомер №2, брак) по имеющимся факторным переменным

**Техническое задание:** требуется проанализировать данные о химических соединениях, выявить зависимости между факторами, построить модель для прогнозирования категорию полученного осадка и дать прогноз. Дать интерпретацию полученным результатам. Сделать выводы.

**Порядок выполнения проекта**

Шаг 1. Выполнить подготовку данных к анализу.

1.1. Загрузить данные в датафрейм.

1.2. Убедиться, что все количественные столбцы имеют числовой тип. Если это не так, выполнить преобразование типа столбца к числовому.

1.3. Проверить данные на пропуски. Если в столбцах или строках очень много пропусков (порог выбрать самостоятельно), такие признаки и объекты для анализа малоинформативны, их следует удалить. Оставшиеся пропуски необходимо заполнить любым подходящим методом (средним, модой, медианой, линейной интерполяцией).

*Результатом* *первого шага* должен быть готовый к анализу набор данных в виде датафрейма.

Шаг 2. Выполнить предварительный анализ данных.

2.1. Вычислить описательные статистики по колонкам (среднее, моду, медиану, стандартное отклонение, квартили).

2.2. Проверить числовые колонки на наличие выбросов, для этого можно использовать диаграмму «ящик с усами» (boxplot). Если выбросов мало, следует их сгладить любым методом (заменой на 1-ый или 99-ый процентиль, среднее, моду, медиану, линейной интерполяцией). Если выбросов много, и они не сконцентрированы в узком диапазоне значений, а рассеяны по широкому диапазону, тогда можно ничего с ними не делать.

2.3. Проверить колонки на нормальность распределения двумя способами:

1) построить гистограмму и сделать предположение о том, являются ли данные нормально распределенными;

2) выполнить статистический тест на нормальность (уровень значимости выбрать самостоятельно) и убедиться, что выдвинутое ранее предположение о нормальности верно (ошибочно).

2.4. Выполнить корреляционный анализ данных с помощью матрицы корреляции. В случае нормальности всех колонок использовать корреляцию Пирсона, в противном случае – ранговые коэффициенты. Матрицу корреляции отобразить с помощью диаграммы «тепловая карта» (heatmap).

2.5. Выбрать целевую переменную – такой признак, который принимает только два значения (метки классов), причем эта метка предположительно может быть предсказана с помощью других признаков. Если такой двоичный признак в данных отсутствует, выбрать такой дискретный или непрерывный числовой признак, который предположительно зависит от других признаков, и сконструировать на его основе целевой признак (добавить новую колонку в датафрейм). Выполнить кодирование целевого признака значениями 0 и 1 (0 соответствует одной метке класса, 1 – другой).

2.6. Используя результаты корреляционного анализа, отобрать факторные признаки – такие, которые предположительно оказывают влияние на целевой и при этом как можно меньше зависят друг от друга.

2.7. Выполнить нормализацию числовых столбцов – факторных признаков путем приведения их значений к диапазону от 0 до 1.

*Результатом* *второго шага* должна быть выдвинутая гипотеза о том, что целевая переменная (класс объекта) зависит от факторных переменных и может быть предсказана с их помощью с применением алгоритмов классификации.

Шаг 3. Выполнить моделирование.

3.1. Разбить данные на обучающую и тестовую выборки (соотношение выбрать самостоятельно, обучающая выборка должна быть больше), предварительно перемешав. Убедиться, что соотношение классов 0 и 1 в обучающей выборке примерно соответствует таковому в тестовой выборке (если в обучающей нулевых классов 40%, то примерно столько же их должно быть и в тестовой). Если это не так, выполнить перемешивание и разбиение снова.

3.2. На обучающей выборке построить несколько моделей, применив различные алгоритмы классификации; использовать в качестве целевой и факторных переменных отобранные ранее столбцы.

3.3. Вычислить показатель качества классификации F-мера на тестовой выборке. Отобрать такую модель, для которой F-мера максимальна. Далее использовать эту модель.

3.4. Используя два любых числовых факторных признака в качестве координат, построить диаграмму рассеяния на тестовой выборке, при этом верно классифицированные единичные объекты обозначить закрашенным кружком, верно классифицированные нулевые объекты – незакрашенным кружком, неверно классифицированные объекты – крестиком.

3.5. Вычислить другие показатели качества классификации: accuracy, precision, recall, построить матрицу ошибок. Дать интерпретацию показателям, подумать, какие из них более информативны и важны для анализа.

*Результатом третьего шага* должны быть выводы о том, насколько хорошо удалось решить задачу классификации (подтвердилась ли выдвинутая на втором шаге гипотеза), какие факторные переменные оказывают наибольшее влияние на целевую, каково объяснение этого, пригодна ли полученная модель для классификации новых объектов.

Шаг 4. Представить отчет по проекту в виде документа Word. Отчет должен включать подробное описание всех выполненных действий с анализом, предположениями о зависимостях в данных и т.д. Обязательно приведение кода на языке Python и результатов его выполнения. Также к отчету должна быть приложена ссылка на блокнот с расчетами в Google Colab (или блокнот Jupyter Notebook, или файл с исходными кодами – если использована иная среда разработки на Python).

Шаг 5. Подготовить презентацию по полученным результатам (в формате PowerPoint). Презентация должна содержать краткое описание хода выполнения проекта (описание данных, их подготовка, предварительный анализ, выдвинутые гипотезы, моделирование) и выводы, сделанные по итогам анализа (подтвердились ли гипотезы?). Количество слайдов – не более 20.

В предоставленном файле **train.csv** описаны параметры уже проведенных реакций и полученных результатов.

**Данные:**

* **product —** результат реакции (***целевая переменная***):  
  0 – смесь веществ (брак),   
  1 – изомер продукта №1,   
  2 – изомер продукта №2.
* **var\_1 —** вес реагента №1 в граммах.
* **var\_2 —** форма реагента №1:  
  0 – порошок,  
  1 – крупные гранулы.
* **var\_3 —**вес реагента №2 в граммах.
* **var\_4 —**форма реагента №2:  
  0 – порошок,  
  1 – крупные гранулы.
* **var\_5 —**отклонение от температуры, указанной в инструкции (87,5**°**С).
* **var\_6 —**материал реактора:  
  0 – стекло,  
  1 – металл.
* **var\_7 —**отклонение от pH, указанного в инструкции (5,8).
* **var\_8 —**использование механического перемешивания:  
  0 – не использовалось,   
  1 – использовалось.
* **var\_9 —**вес катализатора реакции в граммах.
* **var\_10 —**очередность загрузки реагентов в раствор:  
  0 – первым загружен реагент №1,  
  1 – первым загружен реагент №2.
* **var\_11 —**время реакции в минутах.
* **var\_12 —**вес полученного осадка в граммах.

В файле **test.csv** представлена тестовая выборка (параметры реакции без результата). В качестве решения необходимо представить прогноз категории осадка по каждой строчке файла **test.csv.**

Файл **target\_test\_example.csv –** шаблон файл с прогнозом. Именно в таком формате необходимо представить прогноз для автоматизированной проверки точности прогнозирования. Точность прогноза будет оцениваться по метрике «макро-точности»: **средняя точность по каждому классу**.