Санкт-Петербургский политехнический университет Петра Великого

Институт физики, нанотехнологий и телекоммуникаций

# Машинное обучение.Финал

Всероссийская олимпиада "Я - профессионал"

Выполнил: Вылегжанин Евгений Владимирович

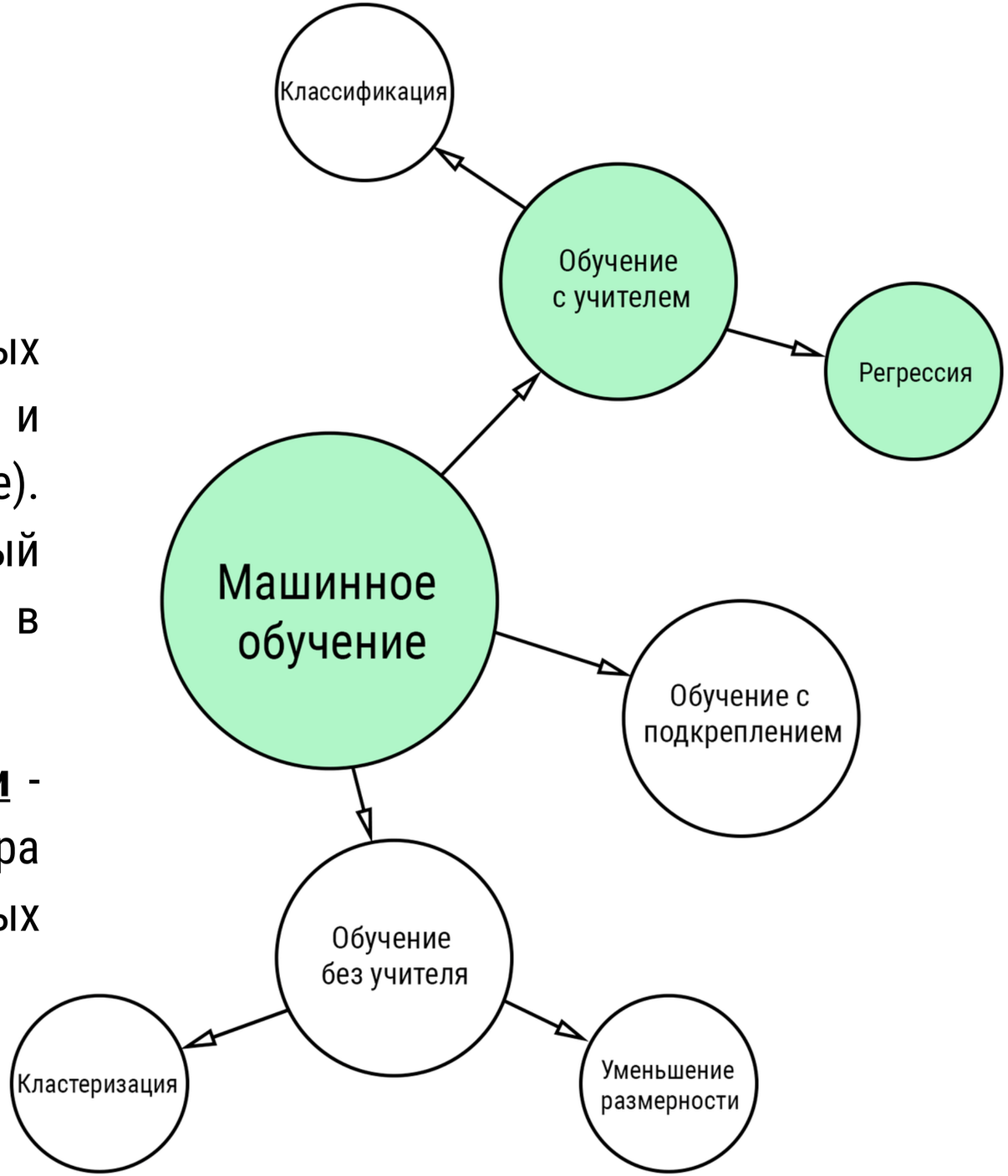
# Содержание

1. Формулировка задачи в терминах машинного обучения
2. Исследование тренировочного и тестового наборов данных
3. Изучение тематической литературы, поиск возможных вариантов решения задачи
4. Список выбранных вариантов решения задач
5. Описание лучшего и наиболее перспективных решений
6. Итоговый слайд со всеми решениями и получившейся метрикой
7. Варианты улучшения решения

# Формулировка задачи

Краткое условие задачи:
Дана информация о 10% скважинных данных (название скажины, две координаты (x, y) и параметр песчанистости в этой точке). Необходимо построить алгоритм, который предсказывал бы значение песчанистости в точках с отсутствовашими значениями.

Данная задача относится к **задаче регрессии** - предсказание численного значения параметра по набору признаков, называемых предсказателями [1].



# Анализ тренировочного набора данных

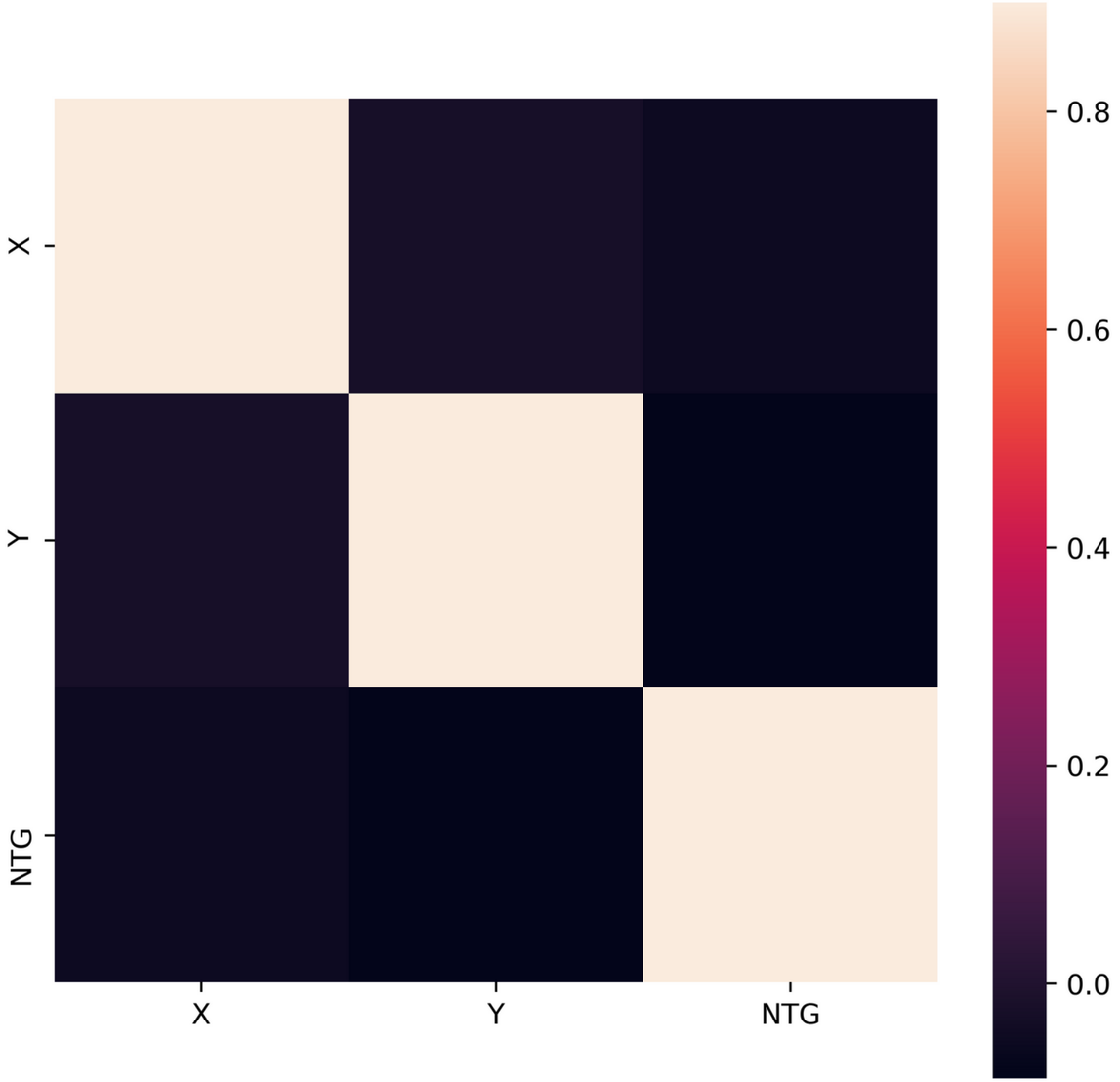
- Признаки: Well, X и Y. Целевая переменная: NTG
- Количество строк в датасете: 138
- Численные признаки: X, Y, NTG
- Категориальные признаки: Well
- Есть ли пропущенные значения, NaN и тд.: нет
- Типы данных различных переменных: Well (object), X и Y (int64), NTG (float64)
- Диапазоны значений: X - [201,  246], Y - [901, 930], NTG - [0.177, 0.563]

Признак "Well" является уникальным для каждого записи тренировочного и тестового датасетов и не несет в себе никакой предиктивной способности, поэтому при дальнейшей работе он не используется.

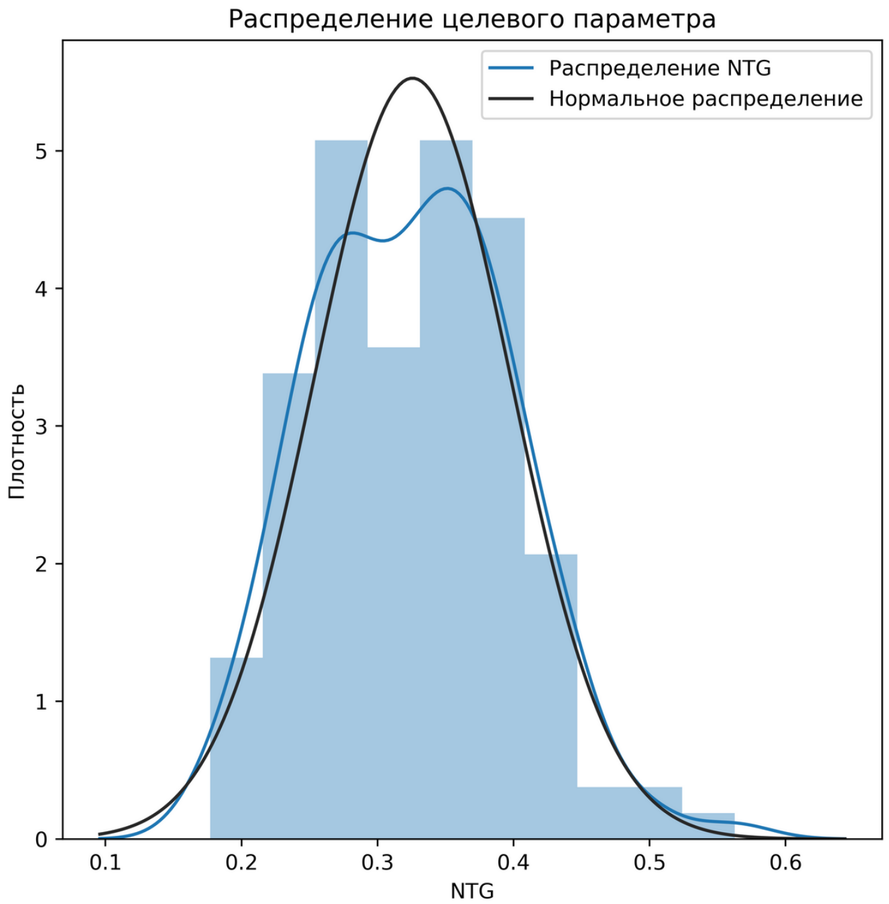
# Корреляция

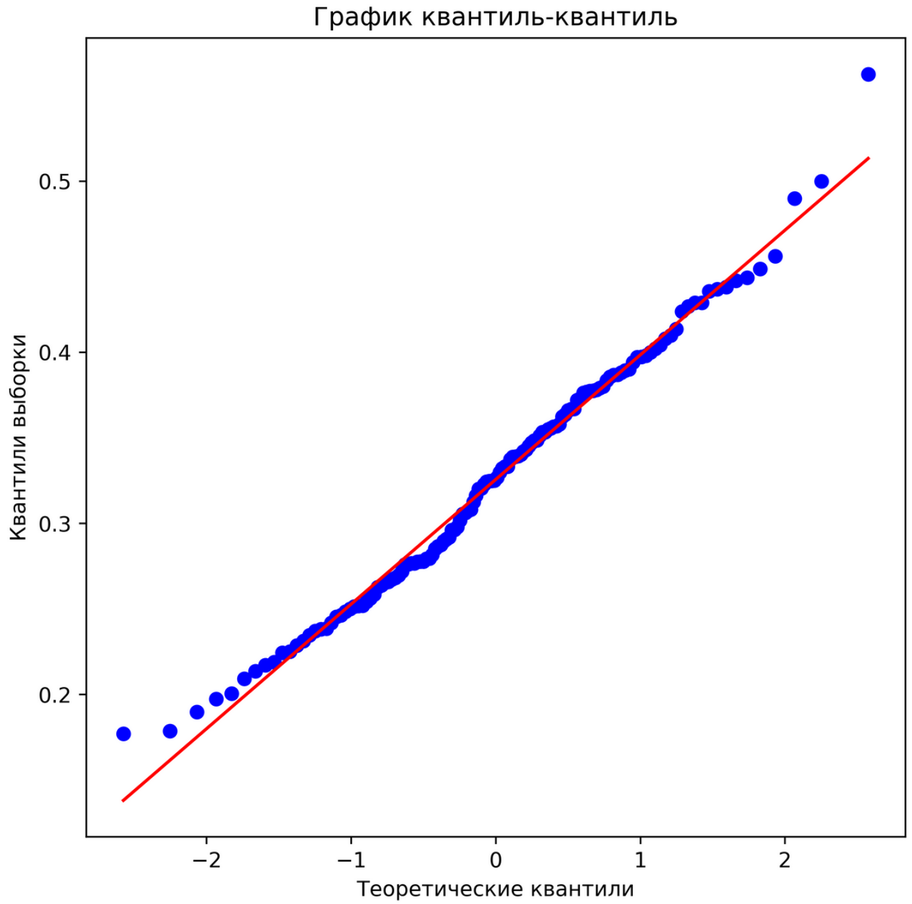
Основная идея данного этапа - проверить насколько целевая переменная зависит от каждого признака; какой вклад каждый признак вносит в предсказание целевой переменной.

Зависимость практически отсутствует. Вывод: для эффективного применения алгоритмов машинного обучения необходимо создать новые признаки на основе существующих.



# Распределение целевой переменной



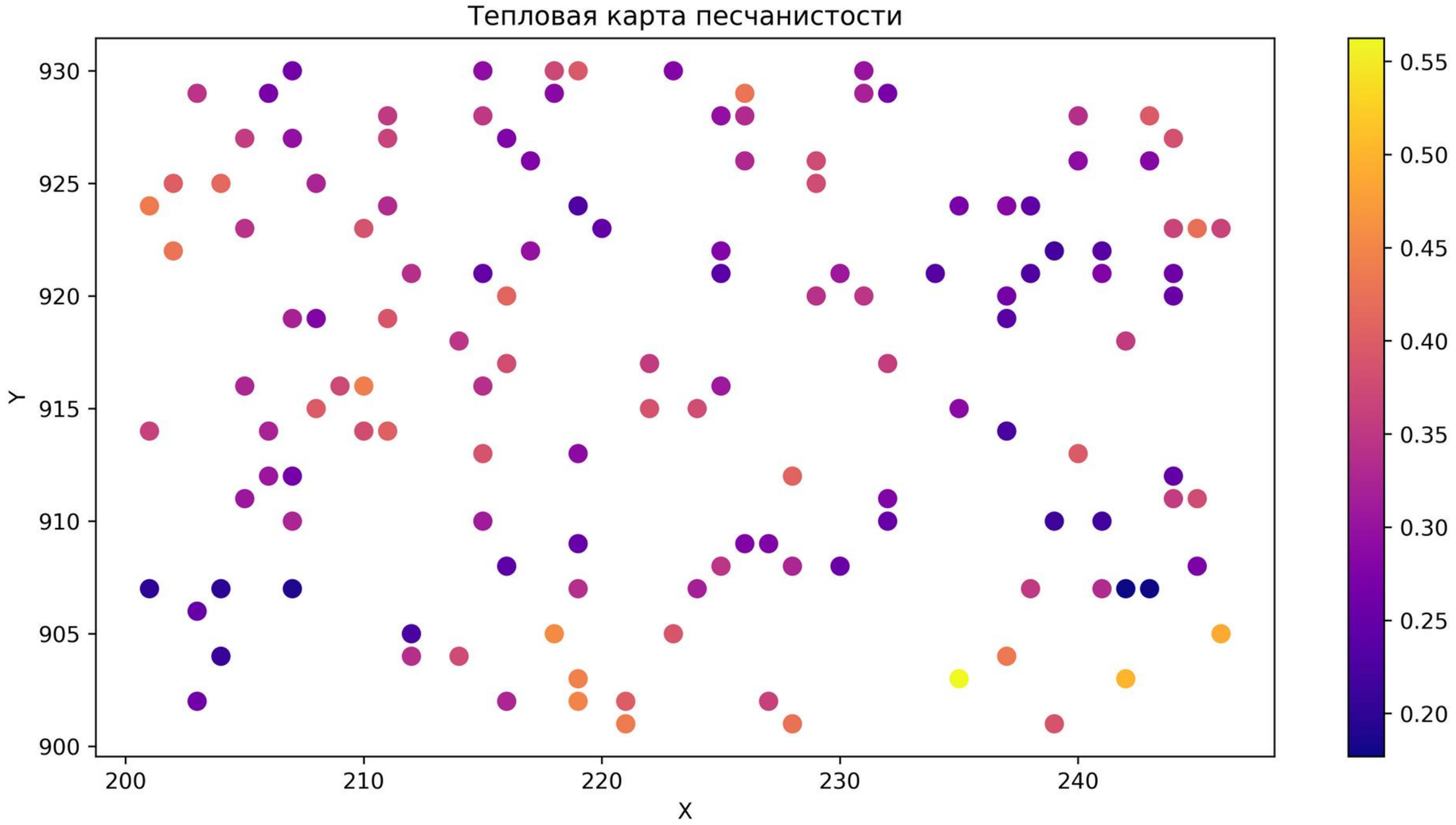

Шапиро-Уилк тест: статистика W = 0.99 (~1), p-value = 0.26 (>0.05) => распределение нормальное

# Анализ тестового набора данных

- Признаки: Well, X и Y
- Количество строк в датасете: 1242
- Численные признаки: X, Y
- Категориальные признаки: Well
- Есть ли пропущенные значения, NaN и тд.: нет
- Типы данных различных переменных: Well (object), X и Y (int64)
- Диапазоны значений: X - [201,  246], Y - [901, 930], NTG - [0.177, 0.563]

# Расположение скважин

# Распределение песчанистости по карте



# Исследование тематической литературы

1) **Geostatistics In Petroleum Geology**
**Olivier Dubrule**

2) Combining Regression Kriging With Machine Learning Mapping for Spatial Variable Estimation

Xiuquan Li, Yile Ao, Shuang Guo, and Liping Zhu

3) Application of machine learning methods to spatial interpolation of environmental variables

Jin Li*, Andrew D. Heap, Anna Potter, James J. Daniell

*Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia*

4) Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania

5) **Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments**

6) **Machine learning and geostatistical approaches for estimating aboveground biomass in Chinese subtropical forests**

7) **Spatial assessment of inland excess water hazard using combined machine learning and geostatistical methods**

8) Thickness, porosity, and permeability prediction: comparative studies and application of the geostatistical modeling in an Oil field

# Варианты решения задачи

1. Кластеризация и усреднение внутри кластеров
2. **Кригинг**
3. **Генерация новых признаков и использование классических методов**
4. Стекинг (с помощью мета-модели или модели второго уровня)
5. Простейший стекинг (усреднение нескольких моделей)
6. **Регрессионный кригинг**

# Кригинг

Создание цифровых геологических моделей осуществляется сегодня на всех этапах жизни месторождения. Подавляющая часть моделей на сегодняшний день является интерполяционными, поскольку распределение пород (и других параметров) в межскважинном пространстве осуществляется интерполяцией (кригингом) скважинных данных [2].

Основная идея кригинга заключается в предсказании значения функции в заданной точке путем вычисления средневзвешенного значения известных значений функции в окрестности точки. Метод тесно связан с регрессионным анализом [7].

# Реализация

Для решения задачи с помощью кригинга использовалось два программных пакета, написанных на Python:

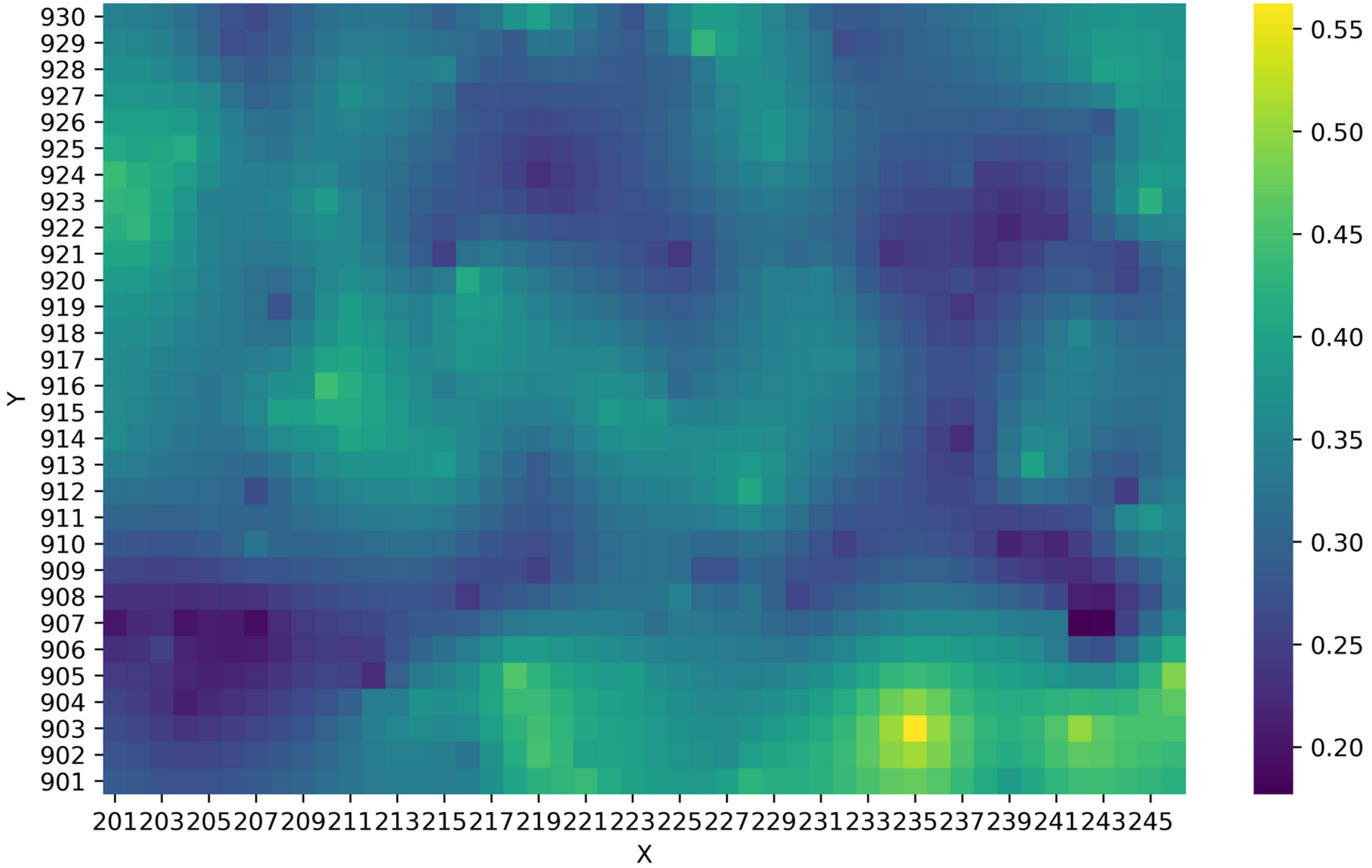
1. **PyKrige** - набор программных инструментов кригинга для Python

С его помощью осуществлялась реализация непосредственно модели кригинга (OrdinaryKriging)

2. **GsTools** - набор программных инструментов геостатистики

Из данного пакета была взята модель ковариации - функция, описывающая степень пространственной зависимости пространственного случайного поля или стохастического процесса

# Тепловая карта лучшего решения



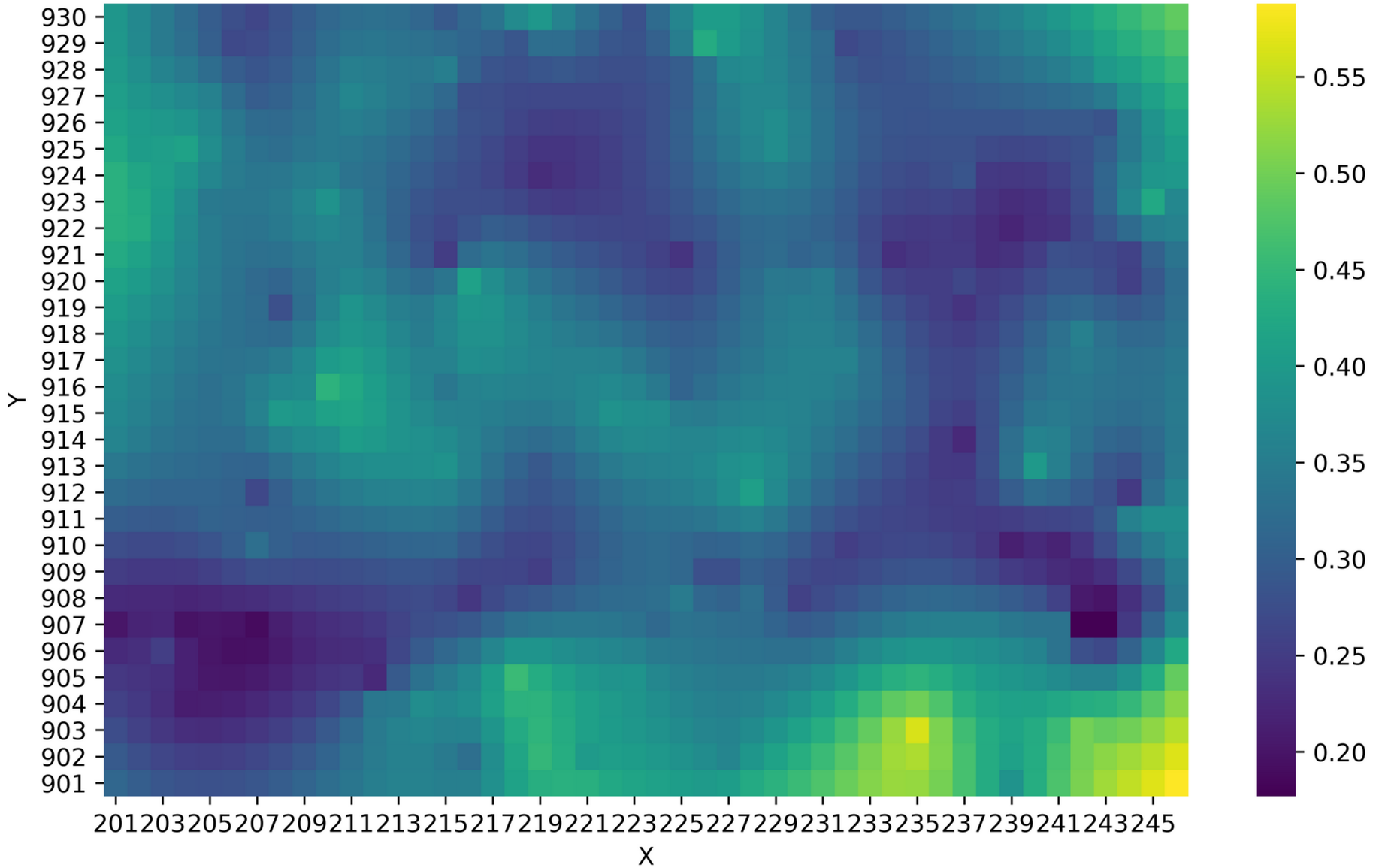
# Создание новых признаков

Как было определено при обработке наборов данных признаки X и Y практически некореллированы с целевой переменной, что делает использование классических методов машинного обучения бесполезным. Для исправления данной ситуации было принято решение создать новые признаки - расстояния от точки на плоскости, соответствующей выбранной скажине, до каждой точки тренировочного датасета. Таким образом получается 138 новых признаков, каждый из которых соответствует расстоянию от текущей точки до одной из 138 точек тренировочного датасета.

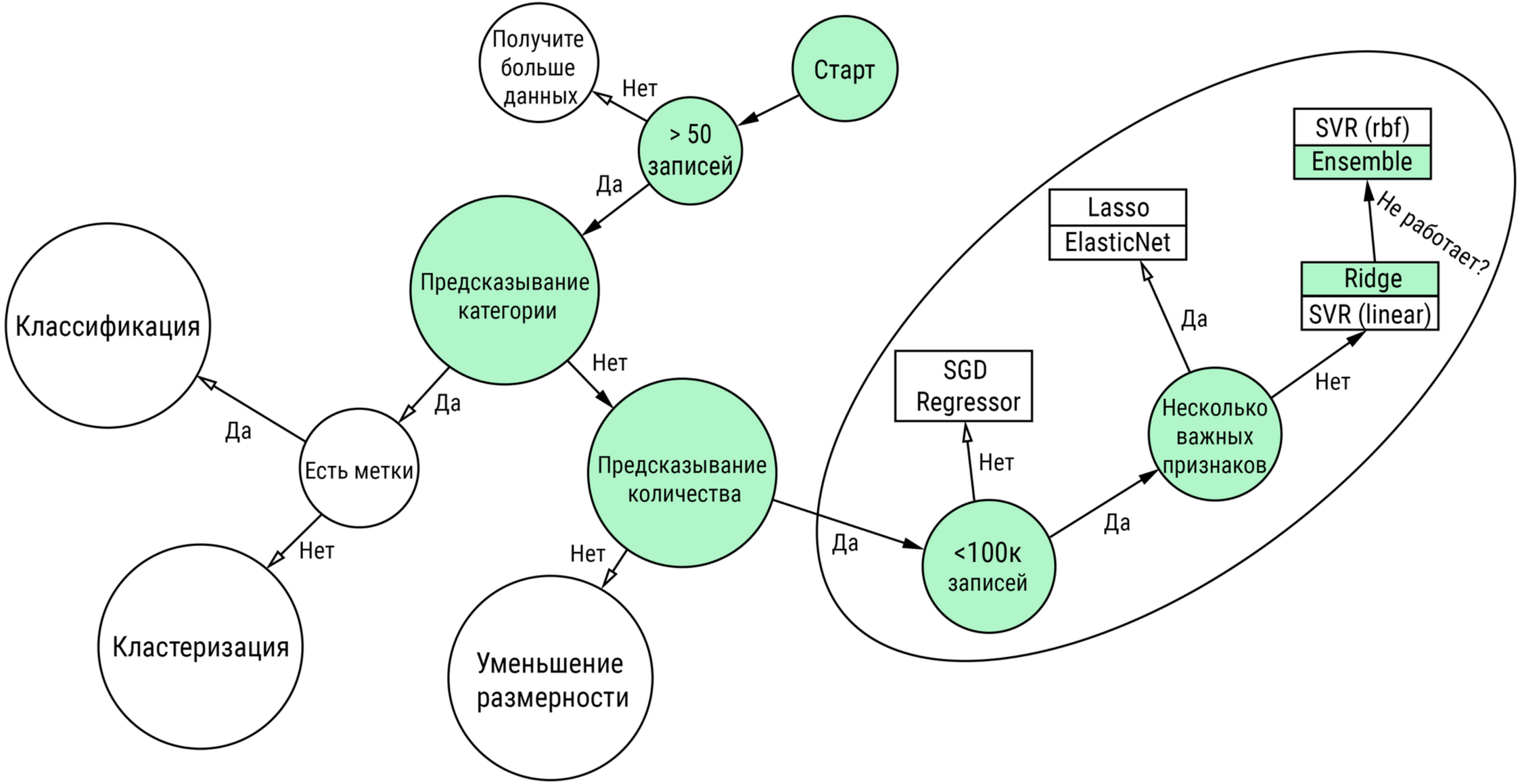
$$R_j^i = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

где $i = 0..137$ – номер нового признака и точки трен. датасета, $j = 0..1379$ – номер записи

# Тепловая карта модели гребневой регрессии



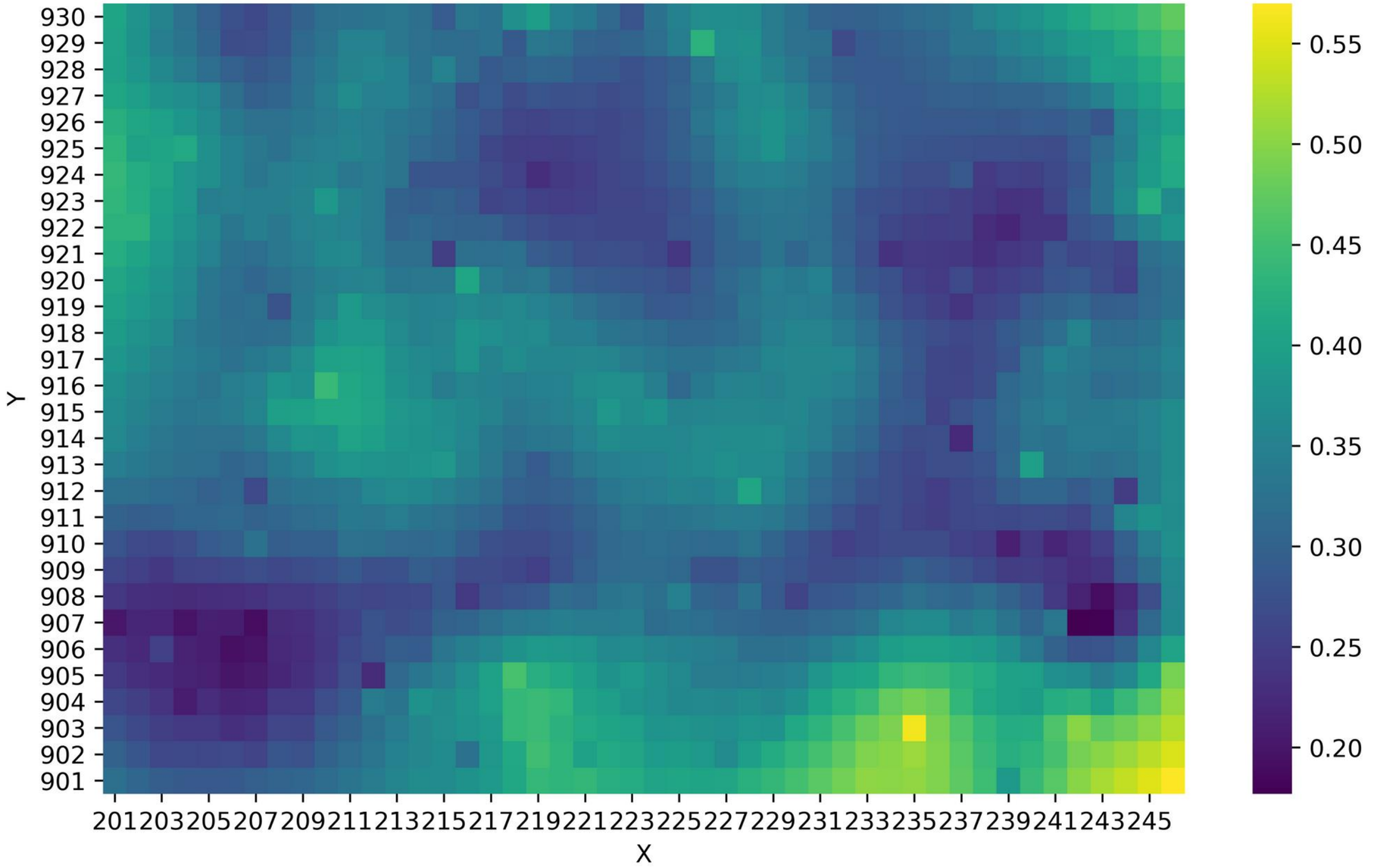
# Выбор алгоритма машинного обучения



# Регрессионный кригинг

Регрессионная модель кригинга представляет собой комбинацию регрессии по методу наименьших квадратов и простого кригинга.
Метод наименьших квадратов (МНК) производит моделирование среднего значения в виде взвешенной суммы независимых переменных (называемой уравнением регрессии), и подразумевается, что ошибки представляют собой случайный некоррелированный шум. Простой кригинг производит моделирование ошибок с помощью модели вариограммы/ковариации, а среднее значение считается постоянным.
Регрессионные модели кригинга позволяют делать более точные прогнозы, чем те, которые могут быть получены при использовании регрессии или кригинга по отдельности [6].

# Тепловая карта модели регрессионного кригинга



# Сводная таблица методов с результатами

| | Баллы | RMSE (CV) | Врем. ресурсы, с |
|---|---|---|---|
| Clustering | 10 | 0.0672 | 0.17 |
| **Kriging** | **90** | **0.0524** | **0.21** |
| **Ridge** | **60** | **0.0581** | **0.14** |
| **Random Forest** | **60** | **0.0539** | **0.86** |
| Stacking (2 lvls) | 30 | 0.0638 | 27.11 |
| Stacking (average) | 50 | 0.0593 | 27.7 |
| **Regression kriging** | **60** | **0.0574** | **0.26** |

# Варианты улучшения результата

1. Исследование регрессионного кригинга (использование различных моделей регрессии и параметров кригинга) - основное направление улучшения результата
2. Оптимизация параметров всех моделей с более мелкой сеткой параметров
3. Анализ случайных процессов при создании моделей месторождений
4. Добавление новых признаков на основе существующих
5. Развитие стекинга (использование других моделей, включение кригинга в стекинг)
6. Анализ и тестирование моделей на исторических данных
7. Тест остальных алгоритмов машинного обучения (XGBoost, CatBoost и тд)

# Список литературы


1) Aurelien Geron, Hands-On Machine Learning with Scikit_Learn and TensorFlow, 564 p., 2017.
2) К.Е. Закревский, В. Л. Попов, Оценка точности интерполяционных геологических моделей. Экспозиция Нефть Газ, Май 3 (56), 2017.
3) Jin Li, Andrew D. Heap, Anna Potter, James J. Daniel, Application of machine learning methods to spatial interpolation of environmental variables, Environmental Modeling & Software, 26 (2011).
4) Оливье Дюбрель, Геостатика в нефтяной геологии. - Москва-Ижевск: Институт компьютерных исследований, НИЦ "Регулярная и хаотическая динамика", 2009, 256 с.
5) Xiuquan Li, Yile Ao, Shuang Guo, Liping Zhu, Combining Regression Kriging With Machine Learning Mapping for Spatial Variable Estimation, IEEE Geoscience and remote sensing letters, 2019.

Электронные ресурсы:

6) https://pro.arcgis.com/ru/pro-app/latest/help/analysis/geostatistical-analyst/what-is-ebk-regression-prediction-.htm
7) https://intellect.icu/kriging-kak-metod-interpolyatsii-ili-regressiya-na-osnove-gaussovskikh-protsessov-9829