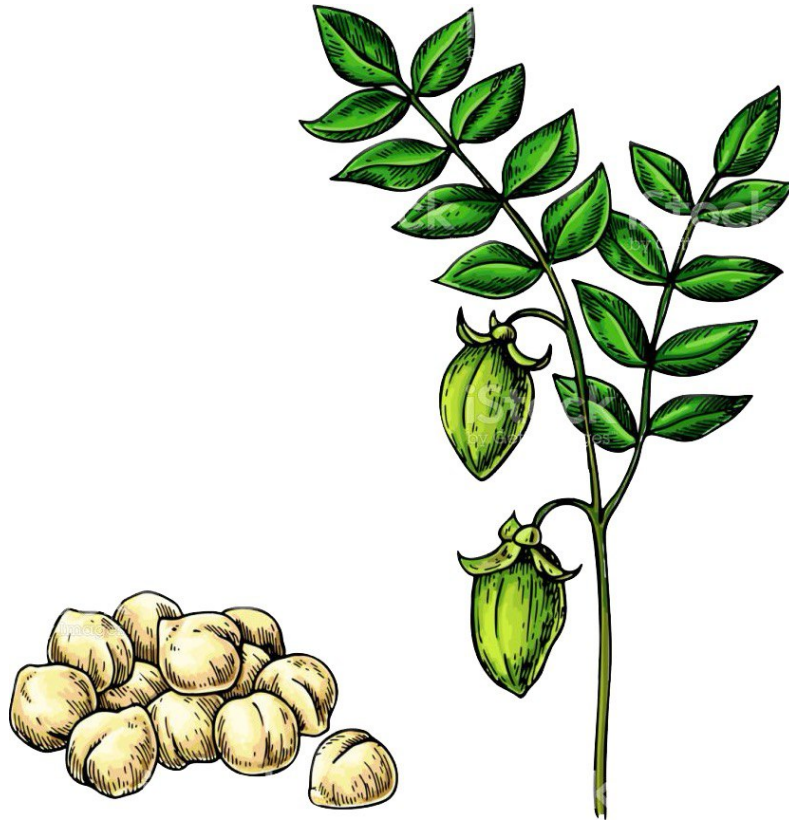
Разработка модели прогнозирования фенотипа растений на основе разреженного разложения искусственных изображений, кодирующих генетические и погодные данные

Студент: Галлямова Нэлли
Научный руководитель: к.б.н. К.Н. Козлов

19.06.2023г.

Введение

- Нут - одно из важных культурных растений, широко используемых как пищевая культура и ценный источник белка. Прогнозирование времени цветения имеет важное значение для улучшения урожайности.
- Искусственные изображения предоставляют удобный способ визуализации и анализа данных, что облегчает процесс извлечения значимых характеристик и позволяет лучше понять взаимосвязи между генетическими и погодными факторами времени цветения нута.



Цель и задачи исследования

- **Цель:** Разработка модели прогнозирования фенотипа растений по генетическим и погодным данным, которые представлены в виде искусственных изображений.
- **Задачи:**
  - Разработка оптимизированного алгоритма кодирования значений факторов в искусственное изображение и методов извлечения характерных черт на основе создания словаря и разреженного кодирования.
  - Создание модели прогнозирования фенотипа растения на основе машины опорных векторов для регрессии.
  - Применение разработанных методов для разработки модели прогнозирования времени цветения нута по имеющемуся набору данных.
  - Выявление факторов, наиболее сильно влияющих на точность модели.

Постановка задачи

Найти функцию F^*, такую что:

$$F^* = \underset{F \in \mathcal{H}}{\operatorname{argmin}} L(y, F(X))$$

$X \in \mathbb{R}^{k \times p}$ - матрица входных данных,
$x_i \in \mathbb{R}^p$ - вектор факторов i-го растения,
$y \in \mathbb{R}^k$ - вектор фенотипов растений,
L - функция потерь,
$\mathcal{H}$ - пространство гипотез.

Набор исходных данных

- Генетические и климатические данные

Генетические данные:

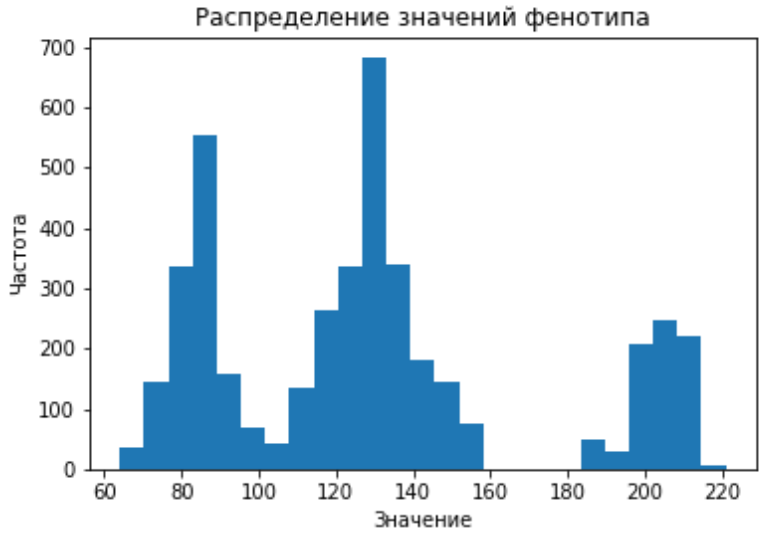
1. Растение и информация о каждом его ОНП

Климатические данные:

1. Значения температур - максимальная и минимальная температуры в течение дня
2. Значение солнечной радиации в течение дня
3. Количество выпавших осадков в течение дня
4. Длительность светового дня

Источники погодных данных:

- 5 дней до дня цветения
- 20 дней после дня цветения

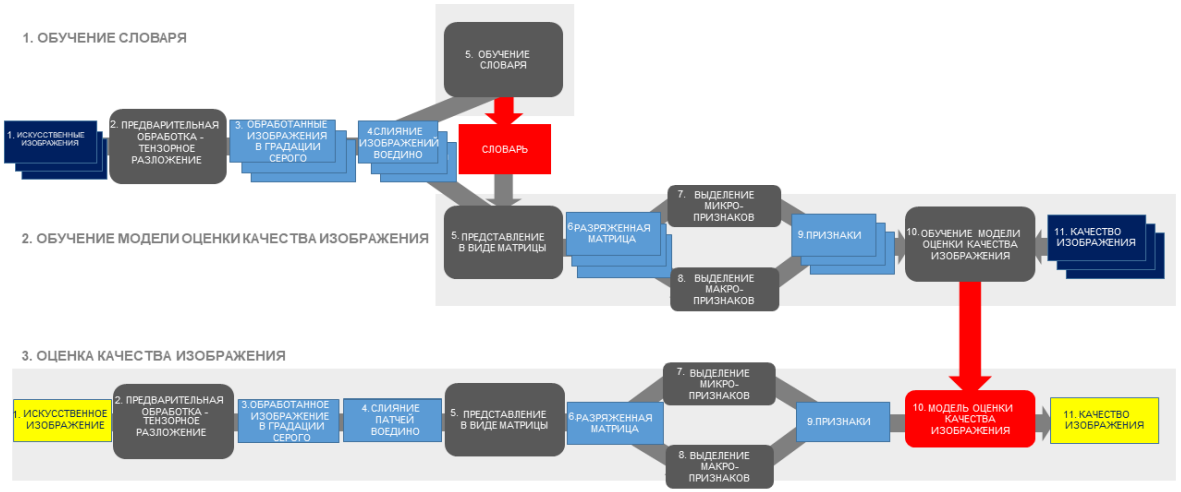


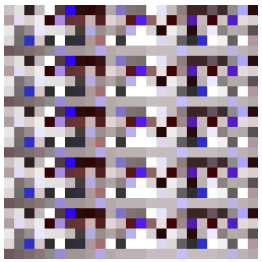
Кодирование данных

- **Кодирование генетической информации**
  - Последовательность ОНП кодируется с использованием битового сдвига.
  - Каждый бит сдвигается влево на количество позиций, равное номеру индекса ОНП в исходной последовательности.
  - Затем применяется операция "установки" бита.
- **Кодирование климатических данных**
  - Преобразование данных путём выполнения целочисленного деления и вычисления остатка от деления.

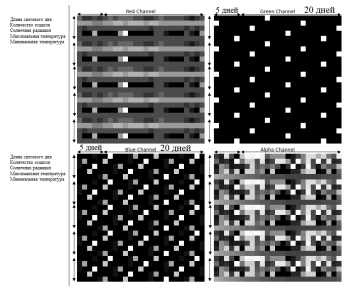
ОНП	Индекс	Значение для r-канала	Значение для g-канала
aa	0	0b0	0b0
aa	1	0b00	0b00
Aa	2	0b100	0b000
AA	3	0b1100	0b1000
aa	4	0b01100	0b01000

Алгоритм оценки качества изображений





Пример изображения

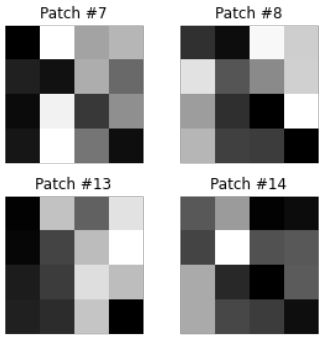


Изображение по каналам в градации серого

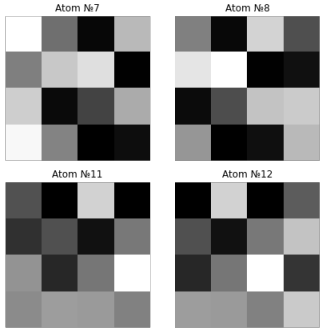
Обучение словаря

- **Основная идея алгоритма**:
  Каждый блок изображения может быть представлен в виде взвешенной суммы изображений-шаблонов (атомов), хранящихся в заранее подготовленном словаре.
- **Задача**: найти наилучший словарь, который представляет входной сигнал как композицию разреженных представлений, решая задачу минимизации:

$$\min_{D,X} \left\{ \|Y - DX\|_F^2 \right\} \forall i, \|x_i\| \leq T_0,$$



Примеры патчей



Примеры атомов словаря

Микропризнаки

- Распределение ненулевых коэффициентов для каждого атома следует логнормальному распределению со следующей функцией плотности вероятности:

$$f(x;\mu;\sigma) = \frac{1}{\sqrt{2\pi}x\sigma} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}$$

$$M[x] = e^{\mu+\frac{\sigma^2}{2}}$$

Вектор микропризнаков для атома:

$$f_i^{\text{mic}} = e^{\mu+\frac{\sigma^2}{2}},$$

где $f^{\text{mic}} = \left[f_1^{mic}, \ldots, f_k^{\text{mic}}\right]$ представляют векторы микропризнаков для всех атомов.

Макропризнаки

- Вероятность появления каждого атома:

$$f_i^{mac} = \frac{n_i}{\sum_{i=1}^{k} n_i}$$

где $f^{mac} = [f_1^{mac}, \ldots, f_k^{mac}]$ представляют векторы макропризнаков для всех атомов, n_i — количество появлений атома i для тестового изображения.

Конечный вектор признаков изображения:

$$f = \left[f^{mic}, f^{mac}\right] \in \mathbb{R}^{2k \times 1}$$

Оптимальная модель: Гиперпараметры и метрики

Оценка качества определяется как $Q = f\left(f^{\text{mic}}, f^{\text{mac}}\right)$
$(f : \mathbb{R}^{2k\times 1} \to \mathbb{R})$

$$f(x) = \sum_{i=1}^{r}(\alpha_i - \alpha_i^*)K(x_i, x) + b$$

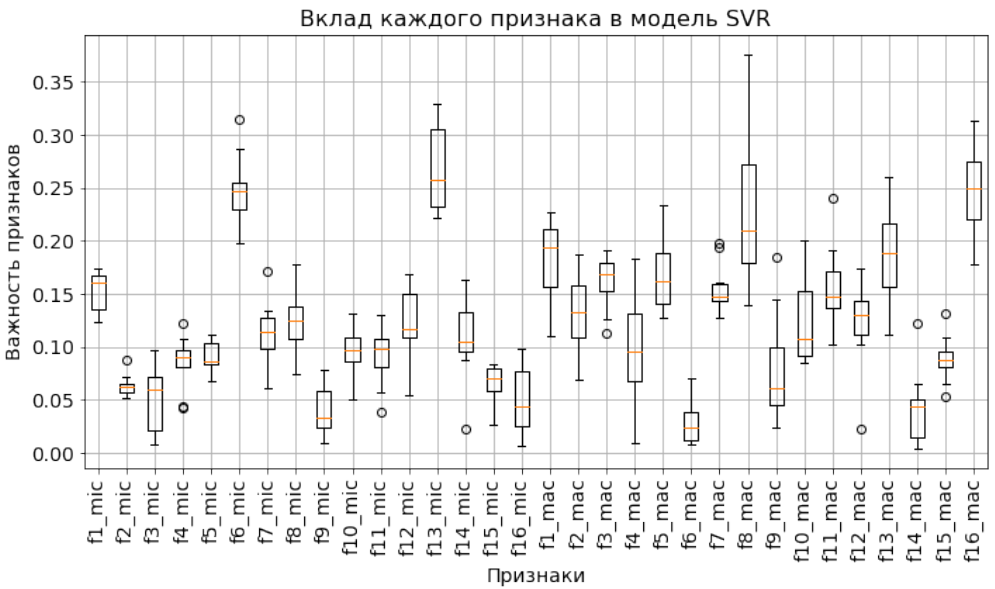
Оптимальные параметры модели:

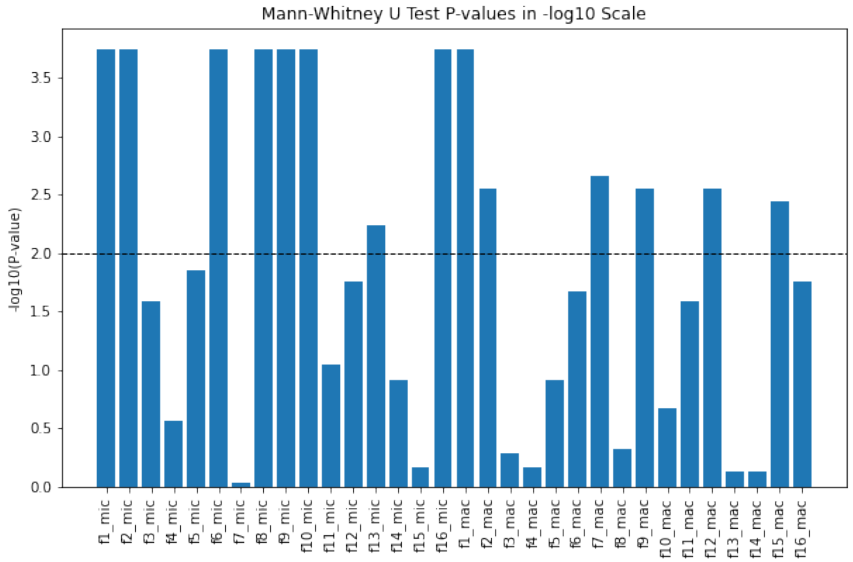
- $C = 1000$
- $\epsilon = 1$
- $\gamma = 1$

Метрики.

- Средняя абсолютная ошибка: **4.36 дня**
- Максимальная ошибка: **14.9 дней**
- Среднеквадратичная ошибка: **5.79 дня**

Оценка вклада признаков в SVR-модель



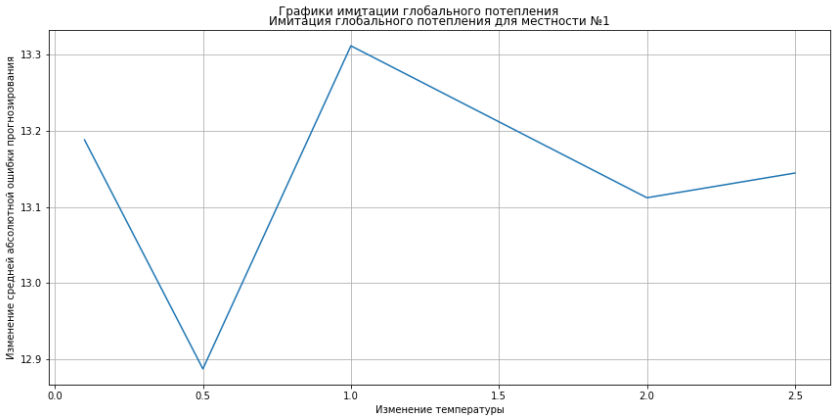


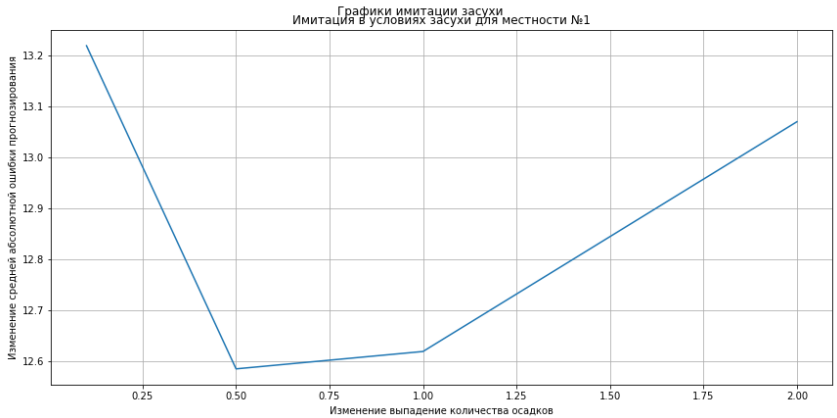
Результаты после исключения менее значимых признаков:

- Средняя абсолютная ошибка: **4.32 дня**
- Максимальная ошибка: **13.99 дней**
- Среднеквадратичная ошибка: **5.45 дня**

Реакция модели в условиях имитации климатических изменений

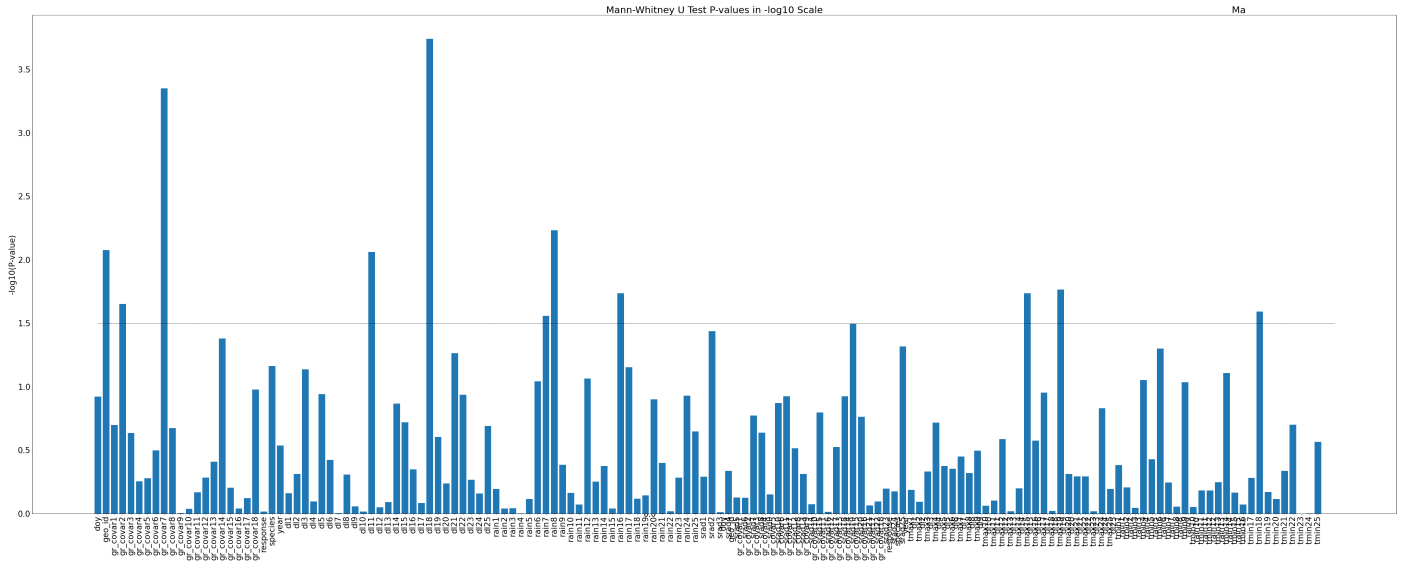
- Имитация: засухи и глобального потепления.
- Моделирование происходило для каждой географической локации отдельно.
- В наборе данных представлены три геолокации.





Несмотря на изменения окружающих условий, модель отражает схожие тенденции изменения фенотипа для всех трех географических локаций, где выращиваются растения. Значения изменений фенотипа колеблются в пределах от 12.5 до 14.5 дней.

Анализ вклада исходных факторов на модель с использованием перестановочного теста и теста Манна-Уитни



- Определены 11 ключевых факторов*, оказывающих наибольшее влияние на модель, дальнейшее изучение которых позволит получить больше информации о том, как они взаимодействуют и влияют на фенотип растений.

*geo_id, gr_covar2, gr_covar7, dl11, dl18, rain7, rain8, srad16, tmax15, tmax19, tmin18

Выводы

- Разработана модель, способная эффективно прогнозировать фенотип растений.
- Анализ вклада факторов позволил улучшить качество модели и снизить ошибку прогнозирования.
- Модель в целом оказалась устойчива к смоделированным изменениям климата.
- Удалось отобрать 11 ключевых факторов, оказывающих наибольшее влияние на модель прогнозирования фенотипа растений.