DATACON 2023

TEAM 7

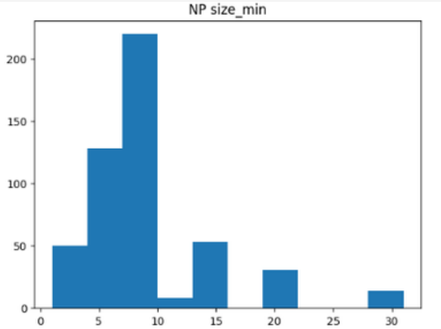
Х.Х. и в продакшн © НТР

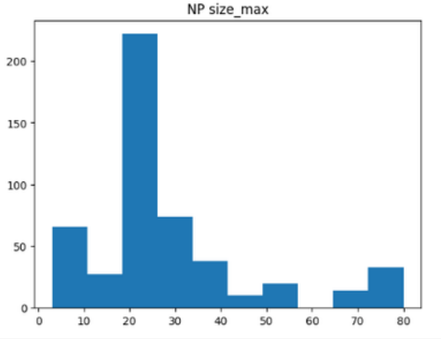
PRESENTED BY:

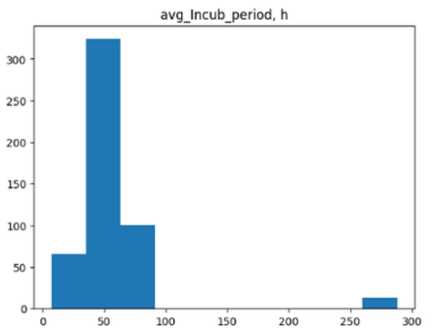
Наталья Гурьева
Вероника Карпушенкова

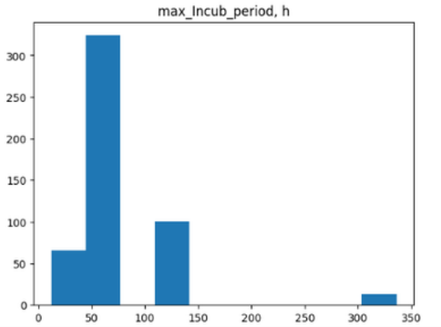
ПРЕДОБРАБОТКА ДАННЫХ

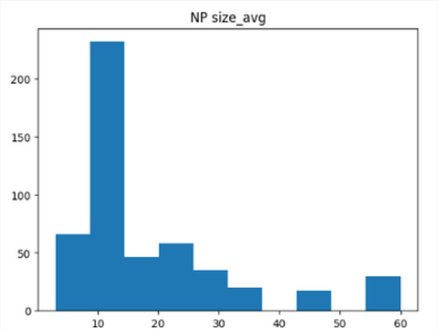
1. убрали ненужные колонки
2. восполнили некоторые значения
3. удалили NaN там, где не смогли восполнить
4. замерджили датасеты
5. создали сабсет с числовым типом данных
6. пронормализовали
7. добавили дескрипторы
8. удалили сильно коррелирующие столбцы

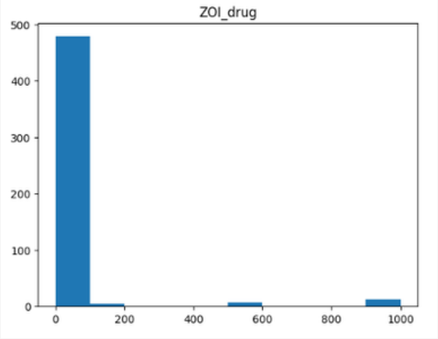

NP size_min

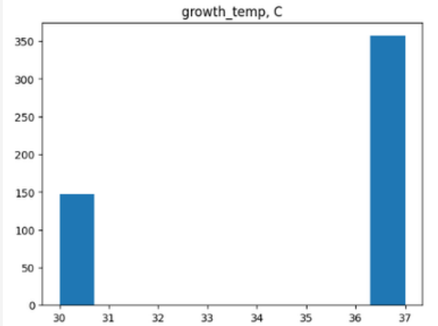

NP size_max

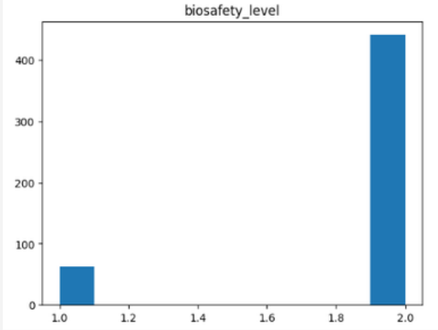

avg_Incub_period, h

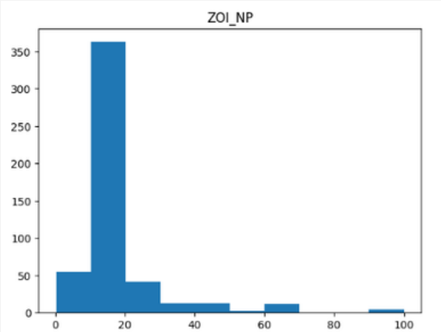

max_Incub_period, h

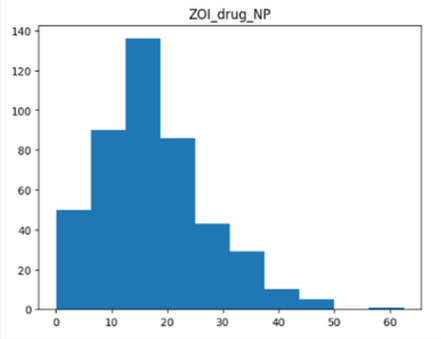

NP size_avg

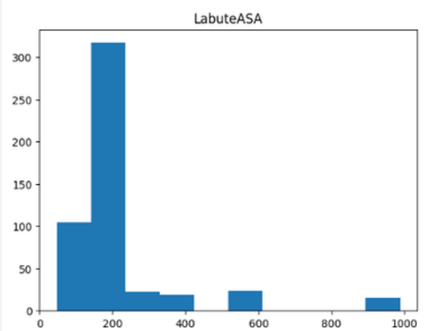

ZOI_drug

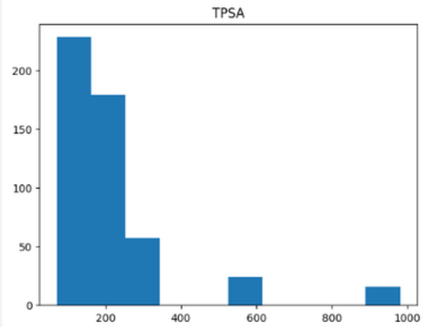

growth_temp, C

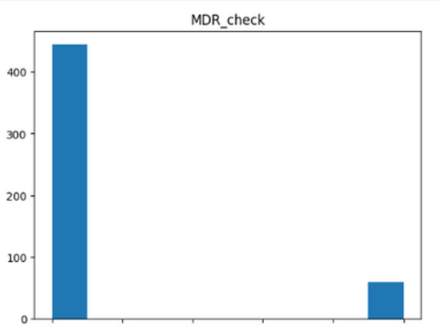

biosafety_level

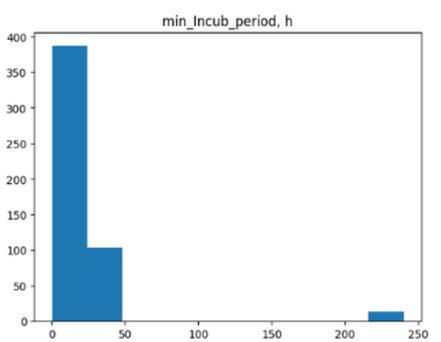

ZOI_NP

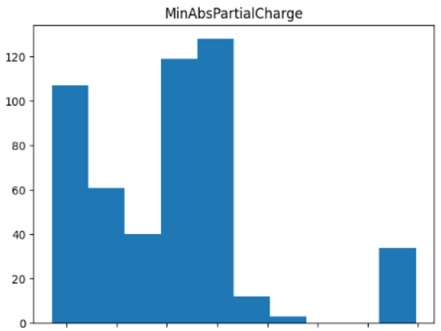

ZOI_drug_NP

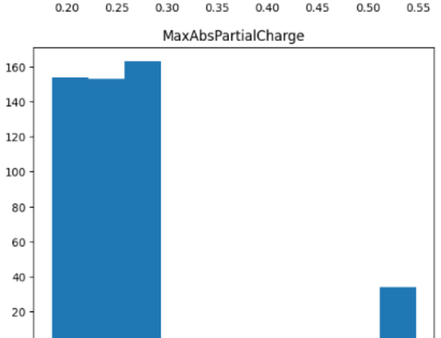

LabuteASA


TPSA


MDR_check

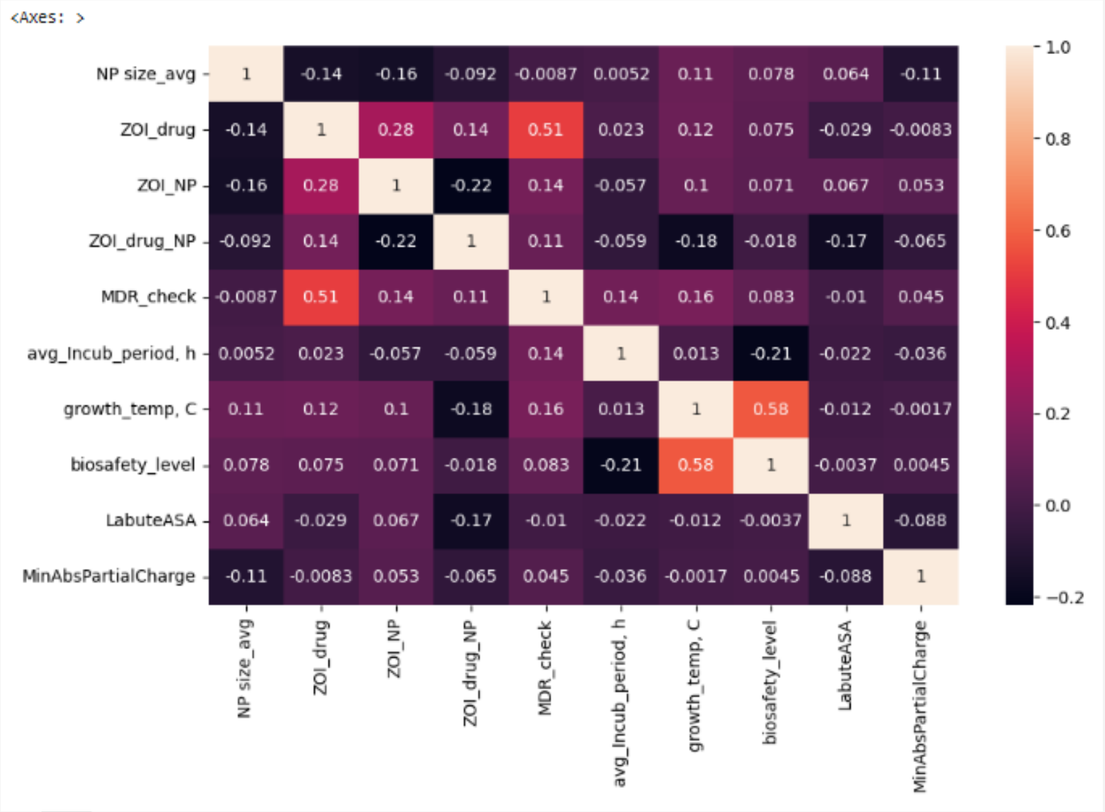

min_Incub_period, h

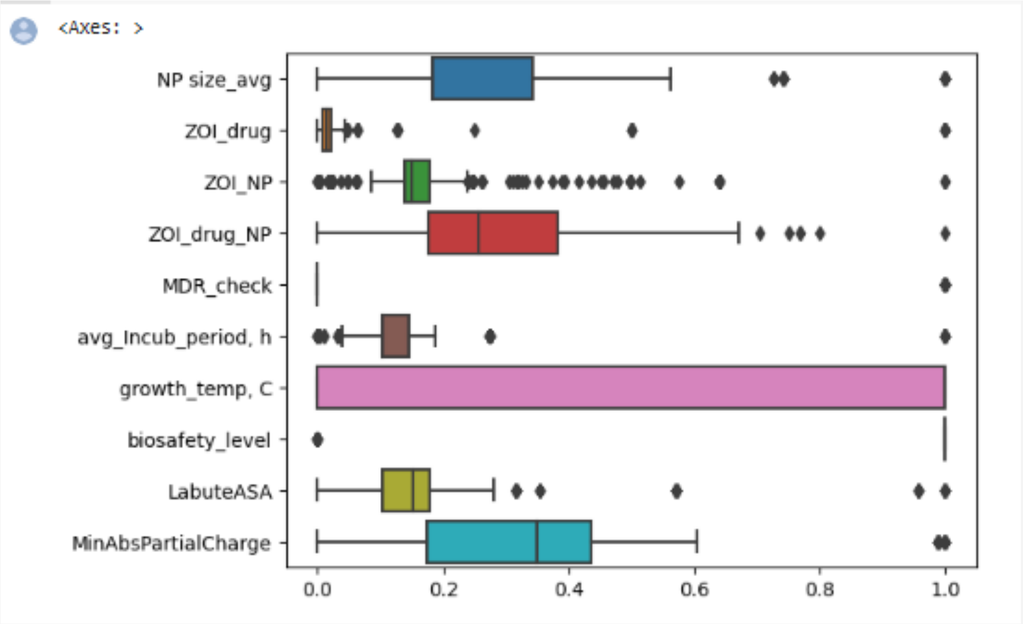

MinAbsPartialCharge

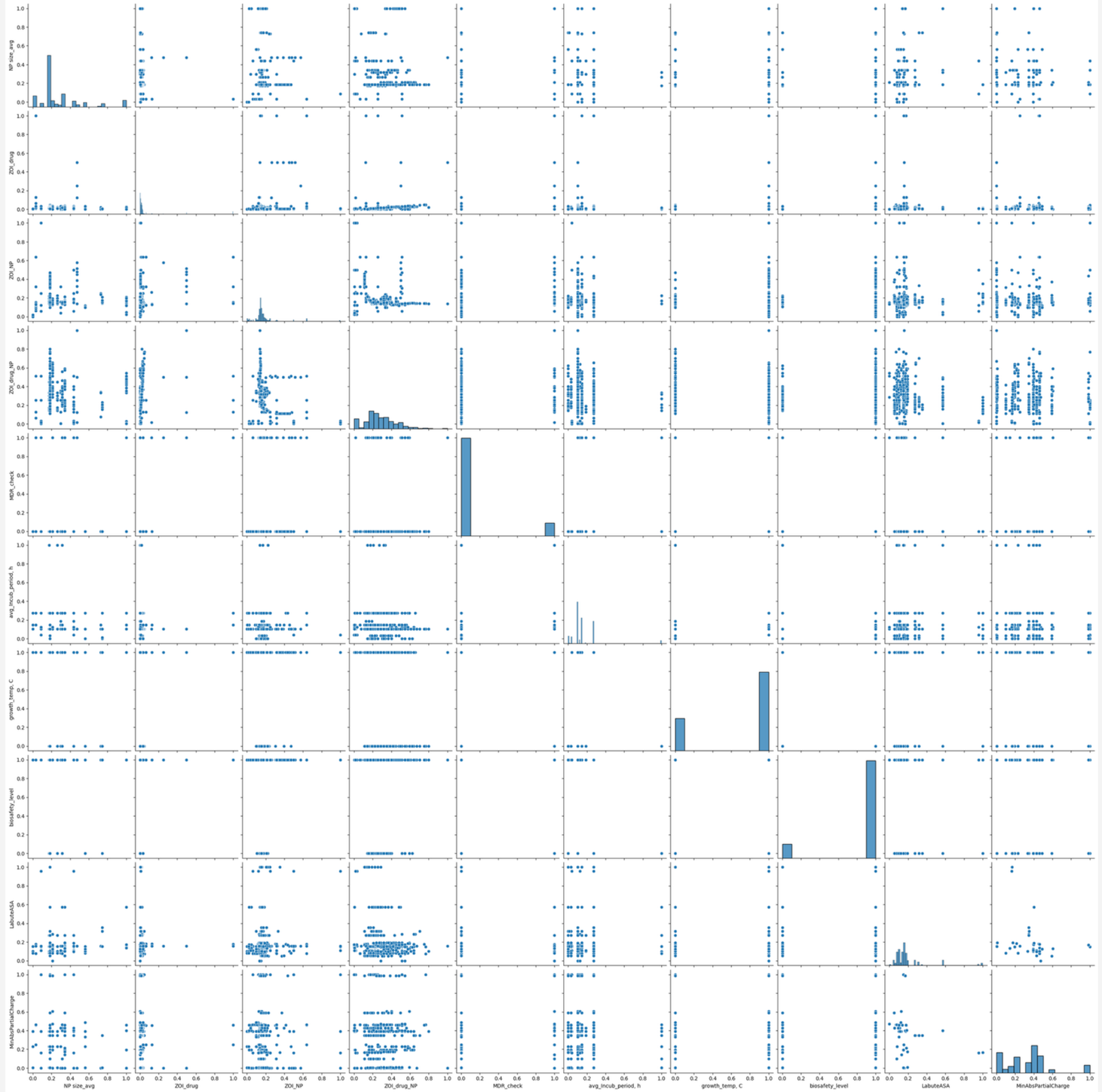

MaxAbsPartialCharge

```
num_subset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   NP size_min           504 non-null    float64
 1   NP size_max           504 non-null    float64
 2   NP size_avg           504 non-null    float64
 3   ZOI_drug              504 non-null    float64
 4   ZOI_NP                504 non-null    float64
 5   ZOI_drug_NP           450 non-null    float64
 6   MDR_check             504 non-null    float64
 7   min_Incub_period, h   504 non-null    float64
 8   avg_Incub_period, h   504 non-null    float64
 9   max_Incub_period, h   504 non-null    float64
 10  growth_temp, C        504 non-null    float64
 11  biosafety_level       504 non-null    float64
 12  LabuteASA             504 non-null    float64
 13  TPSA                  504 non-null    float64
 14  MinAbsPartialCharge   504 non-null    float64
 15  MaxAbsPartialCharge   504 non-null    float64
dtypes: float64(16)
memory usage: 63.1 KB
```






NP size_avg
ZOI_drug
ZOI_NP
ZOI_drug_NP
MDR_check
avg_incub_period_h
growth_temp, C
biosafety_level
LabuteASA
MinAbsPartialCharge


непонел

ВЫБОР ДЕСКРИПТОРОВ

Доказано, что антибактериальная активность зависит от размера, площади и заряда поверхности частицы - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8268496/

Добавили дескрипторы этих свойств: LabuteASA, TPSA, MinAbsPartialCharge, MaxAbsPartialCharge

ОБУЧЕНИЕ МОДЕЛЕЙ

Взяли три модели:

1) линейная регрессия

+: простота, интерпретируемость

2) градиентный бустинг

+: высокая точность, умение работать с различными типами признаков, адаптивность

3) случайные леса

+: устойчивость к выбросам и шумам

ЛИНЕЙНАЯ РЕГРЕССИЯ

Cross-validated scores:

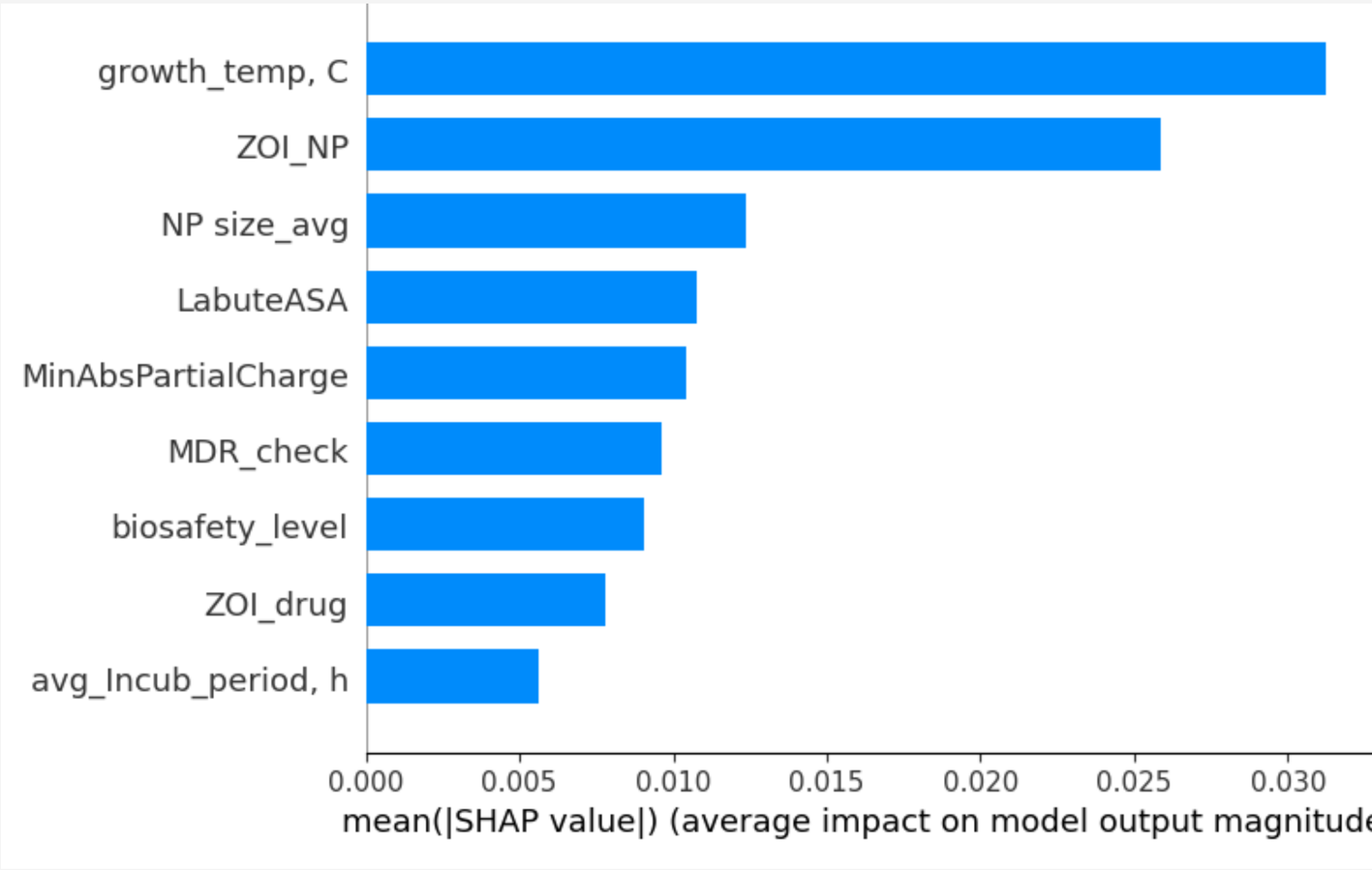
[0.17258129 0.15172355 0.02948637
0.21808306 0.02695485
0.01154977]

Mean absolute error: **0.12**

Mean squared error: **0.02**

Root mean squared error: **0.15**

r2_score: **0.25**



ГРАДИЕНТНЫЙ БУСТИНГ

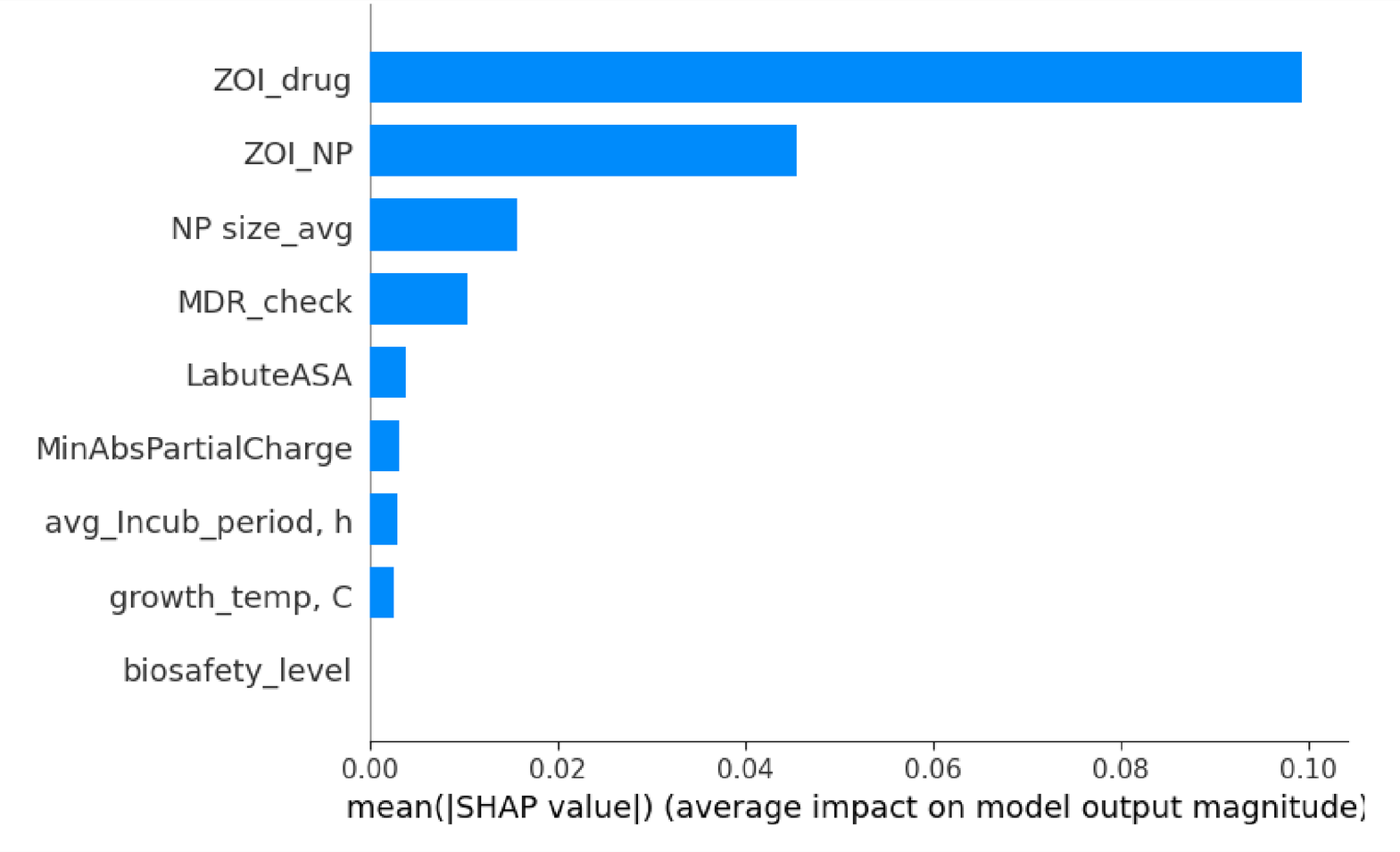
Cross-validated scores:

[0.8005852 0.81062148
0.78668551 0.73552291
0.8993295 0.68110942]

R-squared: **0.787**

MAE: **0.048**

MSE: **0.00597**

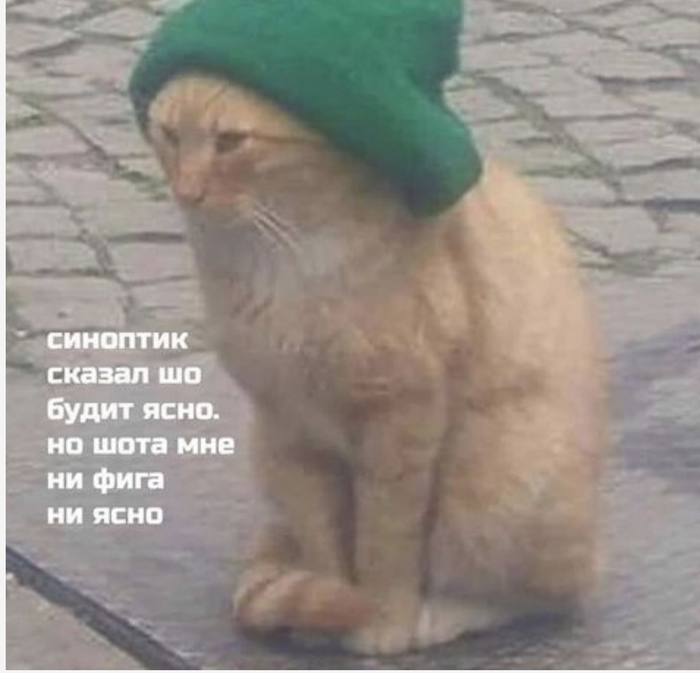


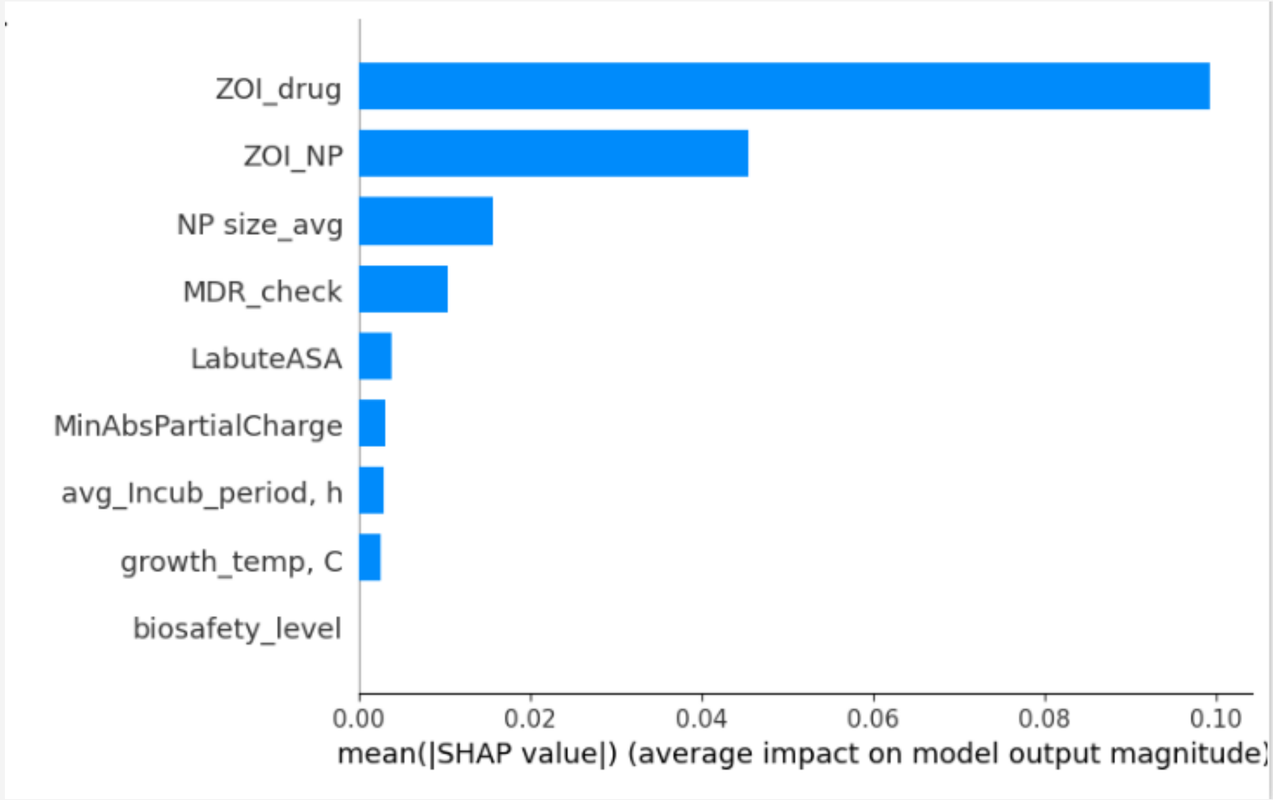
СЛУЧАЙНЫЕ ЛЕСА

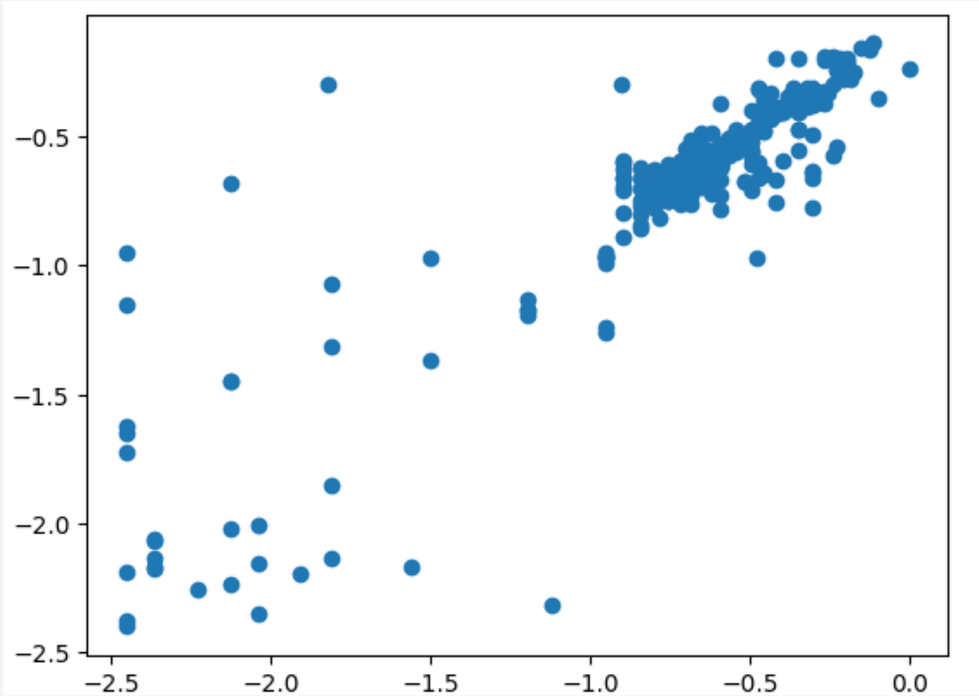
R-squared: **0.794**

MAE: **0.044**

MSE: **0.00580**







DATACON 2023

TEAM 7

PRESENTED BY:

Наталья Гурьева
Вероника Карпушенкова