Сравнительный анализ методов машинного обучения при картографировании массивов открытых песков по спутниковым данным

Полтарин В.С.

Шинкаренко С.С.

Цель:

Рассмотреть основные ансамблевые методы машинного обучения для задач автоматизированного дешифрирования и картографирования опустыненных территорий в аридных ландшафтах на юге России

Актуальность исследования

Проблема опустынивания земель на юго-западе России имеет актуальный статус и активно изучается уже несколько десятков лет. Научный интерес проблемы главным образом связан с решением экономических, экологический показателей территории, а также в разработке автоматизированных средств для мониторинга, прогнозирования и предотвращения подобных ситуаций.

Гиперпараметры, используемые в работе

RandomForestClassifier(random_state=42, max_depth=10, min_samples_leaf=2, min_samples_split=2, n_estimators=100, verbose=3)

ExtraTreesClassifier(max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100, random_state=42, verbose=3)

AdaBoostClassifier(algorithm='SAMME.R', learning_rate=0.1, n_estimators=50, random_state=42)

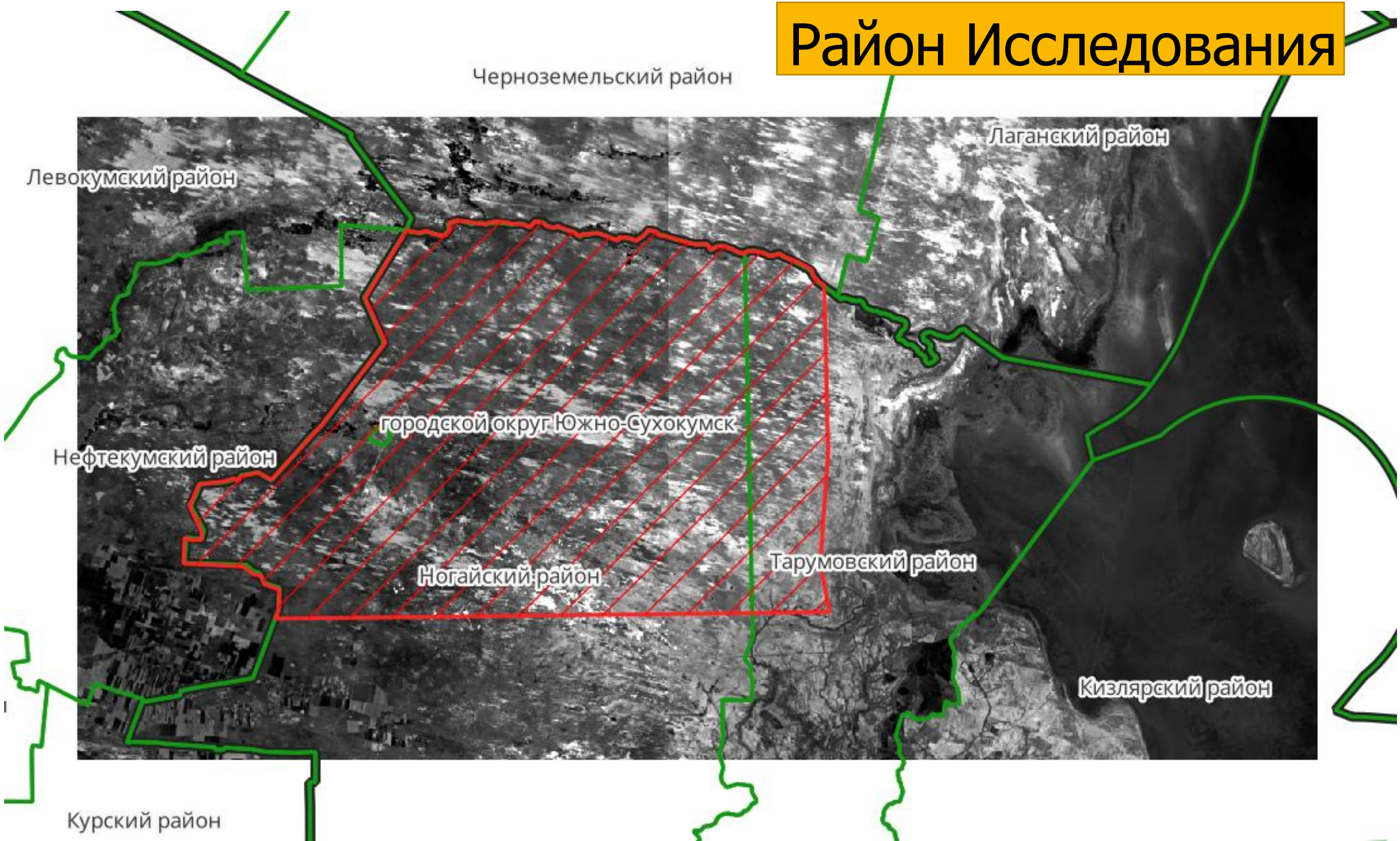
GradientBoostingClassifier(verbose=3, learning_rate=0.01, max_depth=3, max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=200, random_state=42, subsample=0.8)

HistGradientBoostingClassifier(verbose=3, l2_regularization=0.1, learning_rate=0.01, loss='log_loss', max_bins=255, max_depth=3, max_iter=100, max_leaf_nodes=16, min_samples_leaf=1, random_state=42)

BaggingClassifier(verbose=3, base_estimator=None, bootstrap=False, bootstrap_features=True, max_features=0.5, max_samples=0.5, n_estimators=300, random_state=42)

XGBClassifier(verbosity=2, objective='binary:logistic', random_state=42, colsample_bytree=0.8, gamma=0, learning_rate=0.01, max_depth=3, min_child_weight=1, n_estimators=200, reg_alpha=0, reg_lambda=0, subsample=0.8)

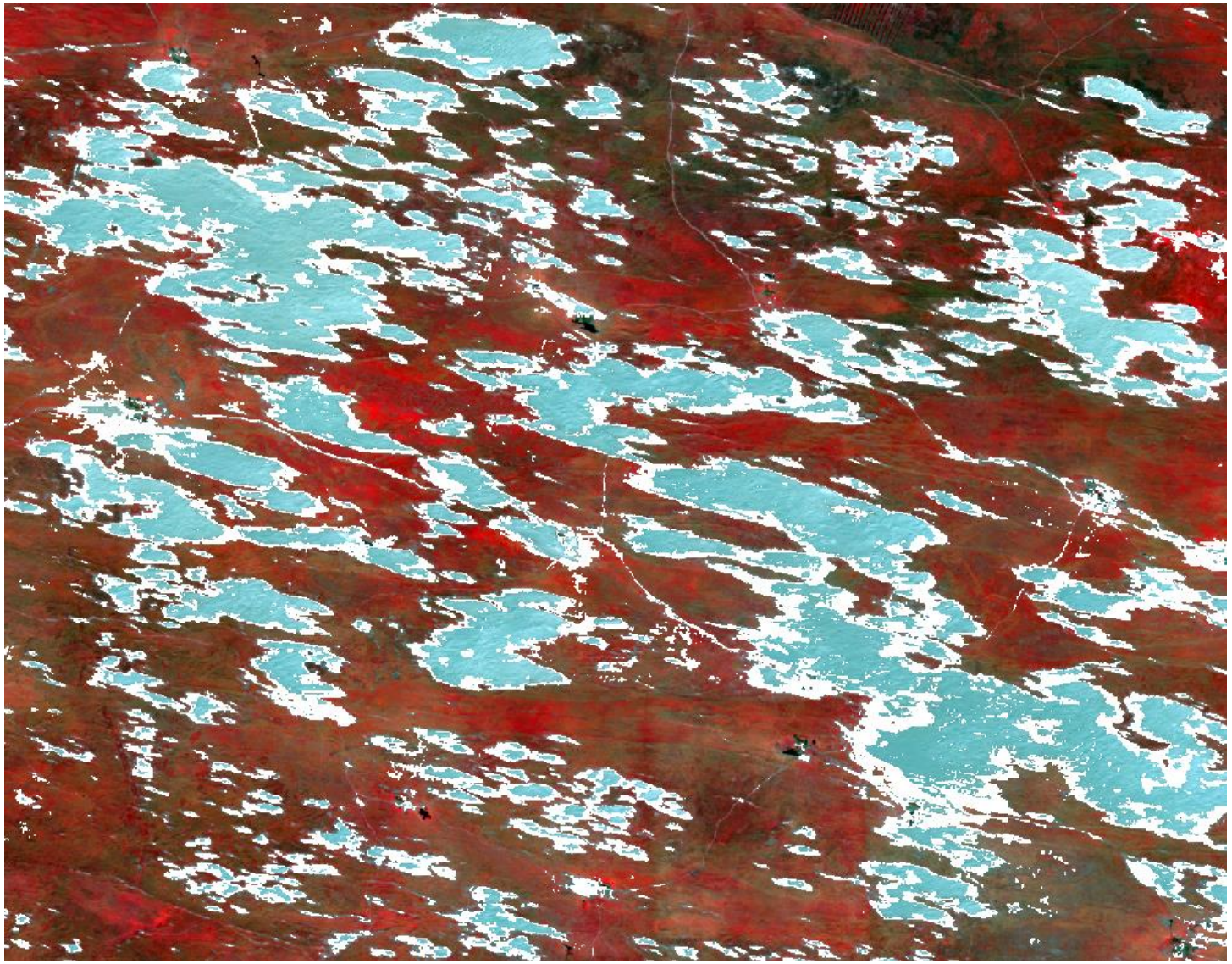
Район Исследования



MIN

MAX

Создание обучающих выборок



МЕТРИКИ ОЦЕНКИ

TP – true positive, FP – false positive, TN – true negative, FN – false negative

Точность (accuracy): Точность измеряет долю правильно классифицированных образцов.
accuracy = (TP + TN) / (TP + TN + FP + FN)

F1-мера (F1-score): F1-мера является гармоническим средним между точностью и полнотой. Эта метрика учитывает как точность, так и полноту, и позволяет найти баланс между ними.
F1 = 2 * (precision * recall) / (precision + recall)

Полнота (recall): Полнота измеряет долю правильно классифицированных положительных образцов среди всех положительных образцов в данных.
recall = TP / (TP + FN)

Точность (precision): Точность измеряет долю правильно классифицированных положительных образцов среди всех образцов, классифицированных как положительные.
precision = TP / (TP + FP)

МЕТРИКИ ОЦЕНКИ

TP	TN	FP	FN
949912	10149400	5178	4757
948225	10149500	5032	6444
949878	10147400	7175	4791
948694	10149600	5009	5975
945427	10150100	4447	9242
948176	10149600	4955	6493
947540	10149700	4850	7129

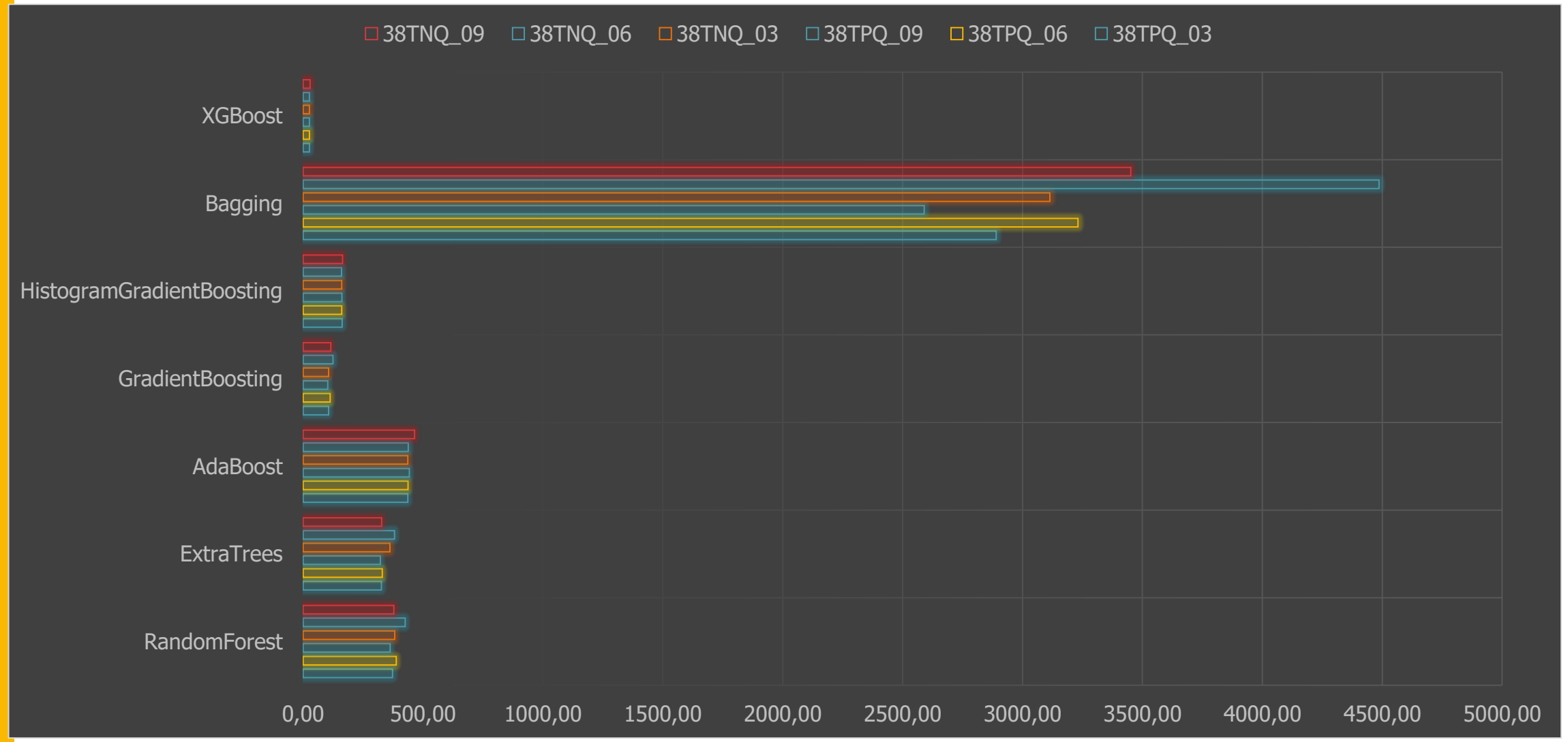
МЕТРИКИ ОЦЕНКИ

Модель	accuracy	precision	recall	f1
RandomForest	0,99910570	0,99910589	0,99910570	0,99910579
ExtraTrees	0,99896699	0,99896640	0,99896699	0,99896664
AdaBoost	0,99892288	0,99892439	0,99892288	0,99892349
GradientBoosting	0,99901127	0,99901087	0,99901127	0,99901105
HistogramGradient Boosting	0,99876778	0,99876615	0,99876778	0,99876638
Bagging	0,99896951	0,99896887	0,99896951	0,99896913
XGBoost	0,99892171	0,99892081	0,99892171	0,99892112

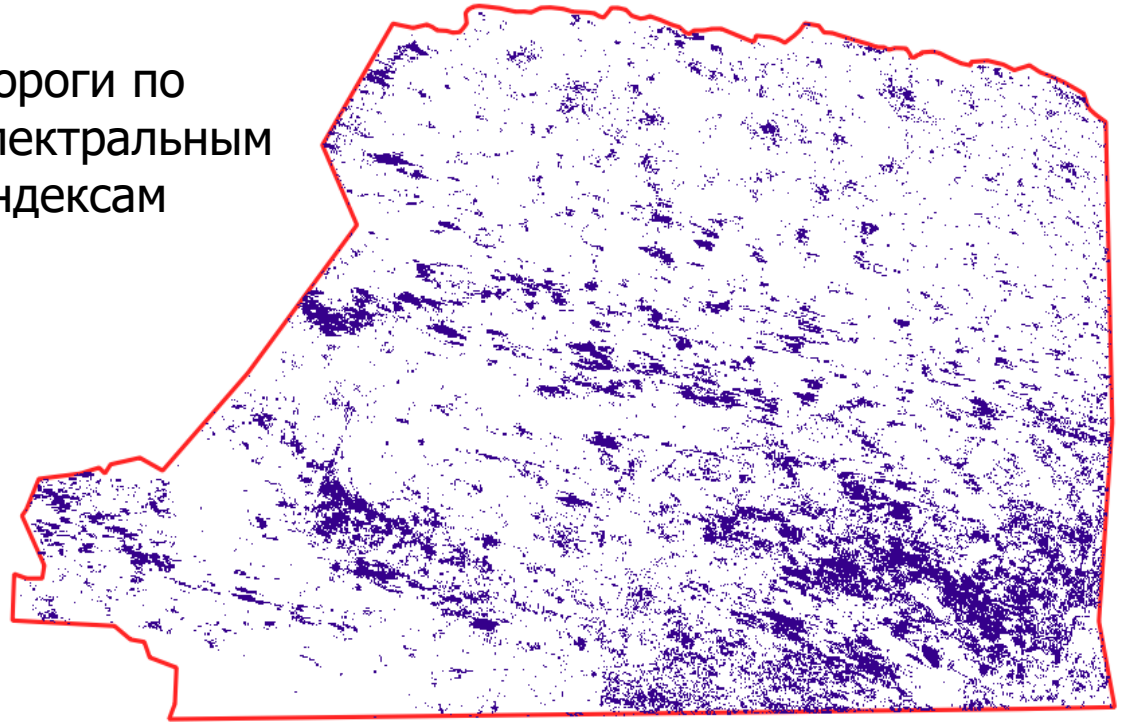
Время обучения

Модель	время подбора гиперпараметров (в мин)	время обучения модели (в мин)
RandomForest	47,03	135,93
ExtraTrees	47,03	23,12
AdaBoost	6,23	41,55
GradientBoosting	1074,27	91,69
HistogramGradient Boosting	4,67	2,90
Bagging	662,29	618,54
XGBoost	22,94	5,73

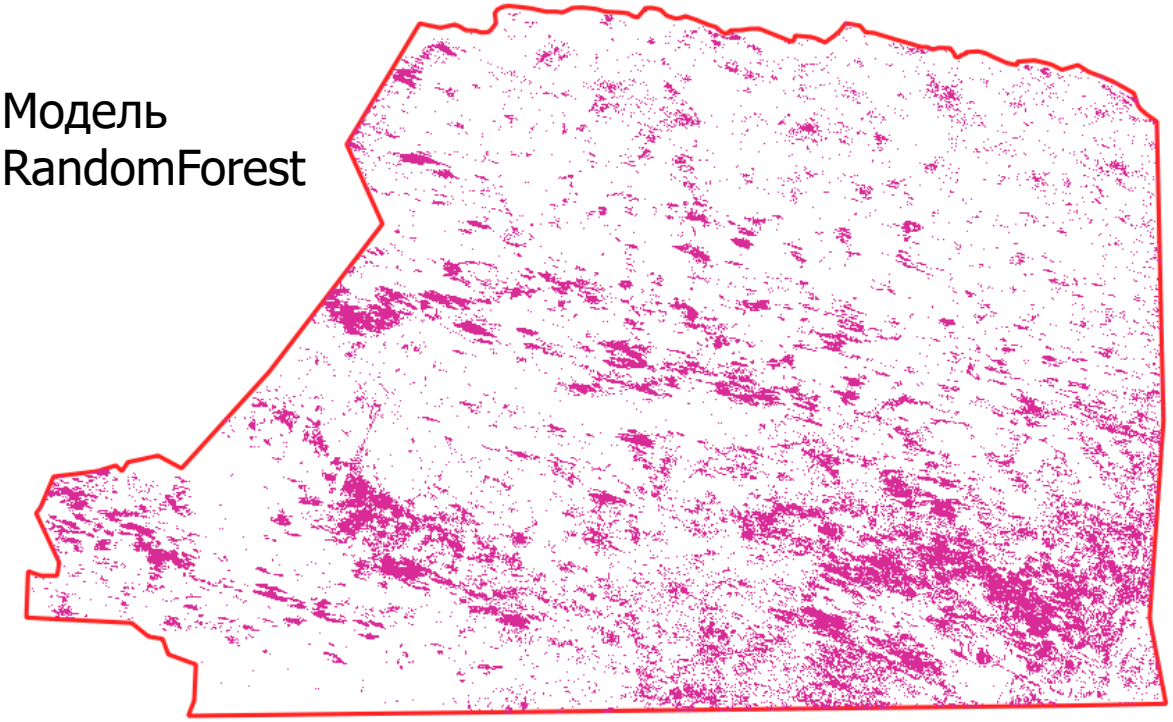
Время предсказания по тайлам (сек.)



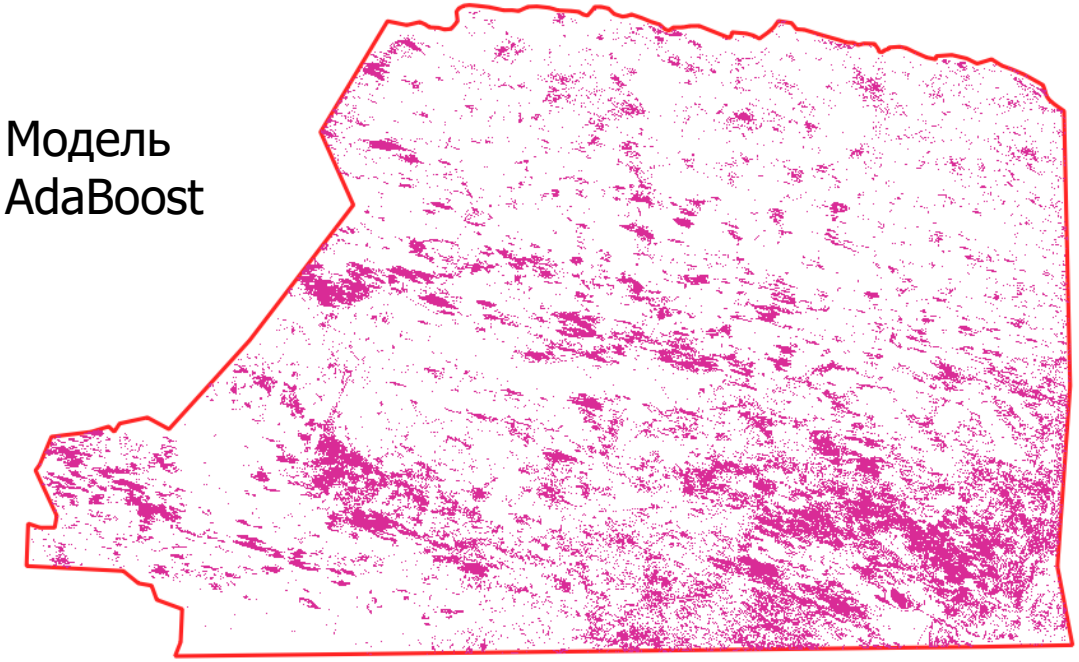
Пороги по спектральным индексам



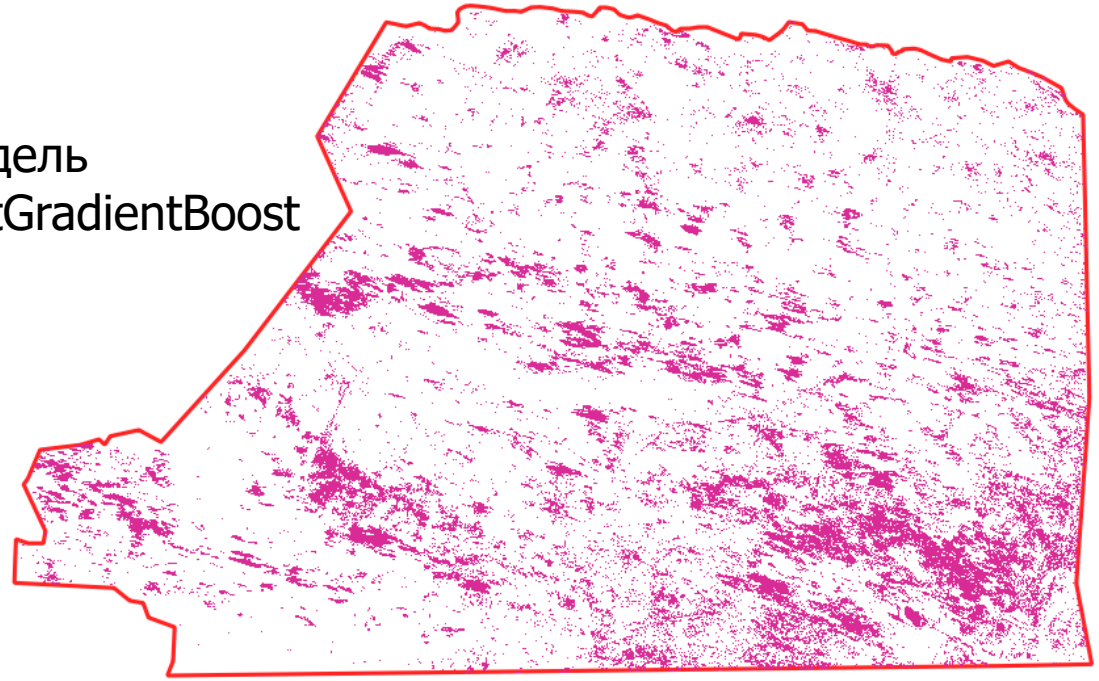
Модель RandomForest



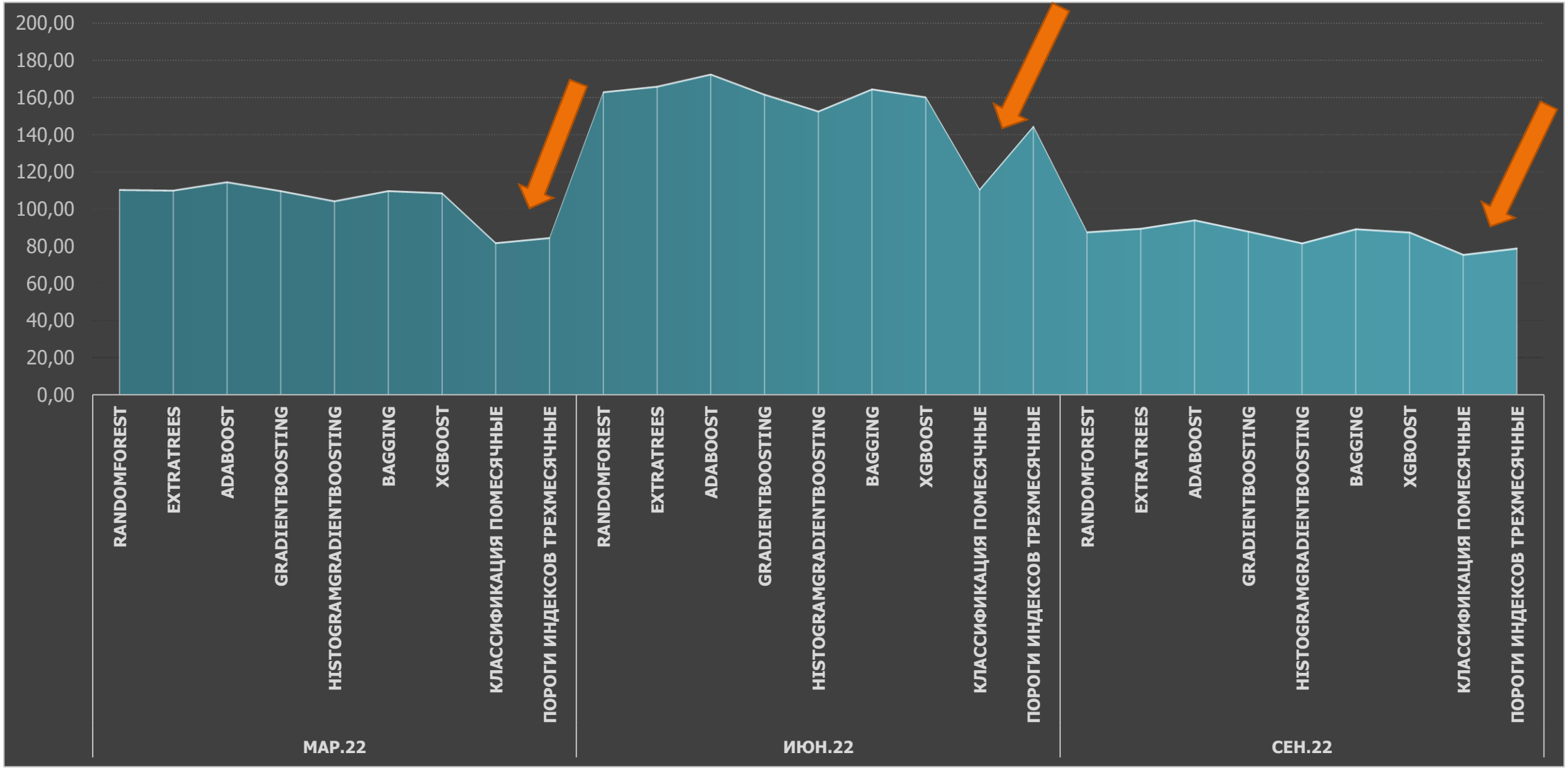
Модель AdaBoost



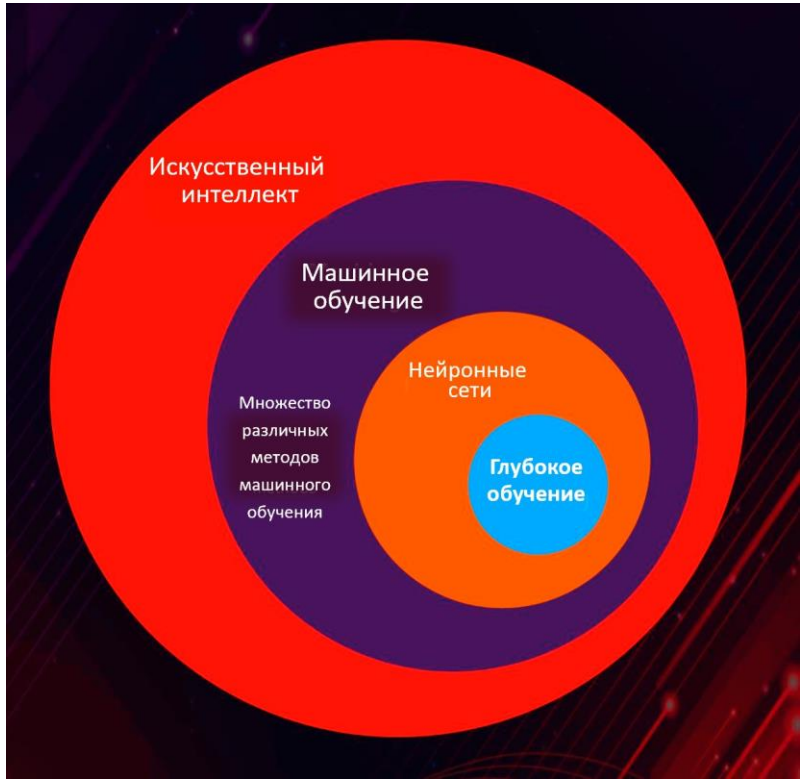
Модель HistGradientBoost



Сравнение площадей, (км кв.)



Перспективы развития – нейронные сети



- Библиотека PyTorch
- Библиотека TensorFlow + Keras
- ArcGIS Deep-learn

Заключение

- Ансамблевые методы, представленные в работе, показали высокую точность классификации открытых песков на основе обучающей выборки по индексам NDVI. Эти методы обеспечивают стабильные результаты и эффективно справляются с проблемой переобучения.

- Подбор гиперпараметров ансамблевых методов является важным шагом для улучшения качества работы модели. Сетка поиска и случайный поиск позволяют выбрать оптимальные значения гиперпараметров, обеспечивая высокую точность классификации.

- Правильный подбор обучающих данных является неотъемлемой частью процесса построения модели машинного обучения и имеет решающее значение для ее эффективности и качества работы.

- Ансамблевые методы могут быть эффективно использованы для решения практических задач дистанционного зондирования Земли, таких как определение площадей открытых песков, в том числе и на основе спектральных индексов.