



ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ ПРИ РЕШЕНИИ ГЕОИНФОРМАЦИОННЫХ ЗАДАЧ

Колесников А.А., Кикин П.М., Комиссарова Е.В., Касьянова Е.Л.

СГУГиТ, г. Новосибирск

«ИнтерКарто/ИнтерГИС 24», г. Петрозаводск, 2018

Машинное обучение, особенности

Формулировка задачи: сбор параметров



Решение задачи: процесс обучения

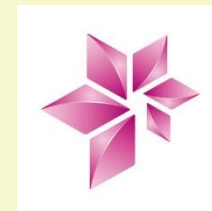


Проверка модели


- ✓ Популярность
- ✓ Высокий порог вхождения
- ✓ Универсальность

Задачи географии, картографии и геоинформатики, решаемые средствами машинным обучением

- ✓ **классификация** – отнесение объекта к одной из категорий на основании его признаков;
- ✓ **регрессия** – прогнозирование одного или нескольких количественных признаков объекта на основании набора прочих его параметров (как количественных, так и качественных),
- ✓ **кластеризация** – разбиение множества объектов на группы на основании признаков этих объектов;
- ✓ **детекция аномалий** – поиск объектов, сильно отличающихся от всех остальных в выборке либо от какой-то группы объектов.



Sharma, Diksha & Kumar, Neeraj. (2017). A Review on Machine Learning Algorithms, Tasks and Applications. 6. 2278-1323.

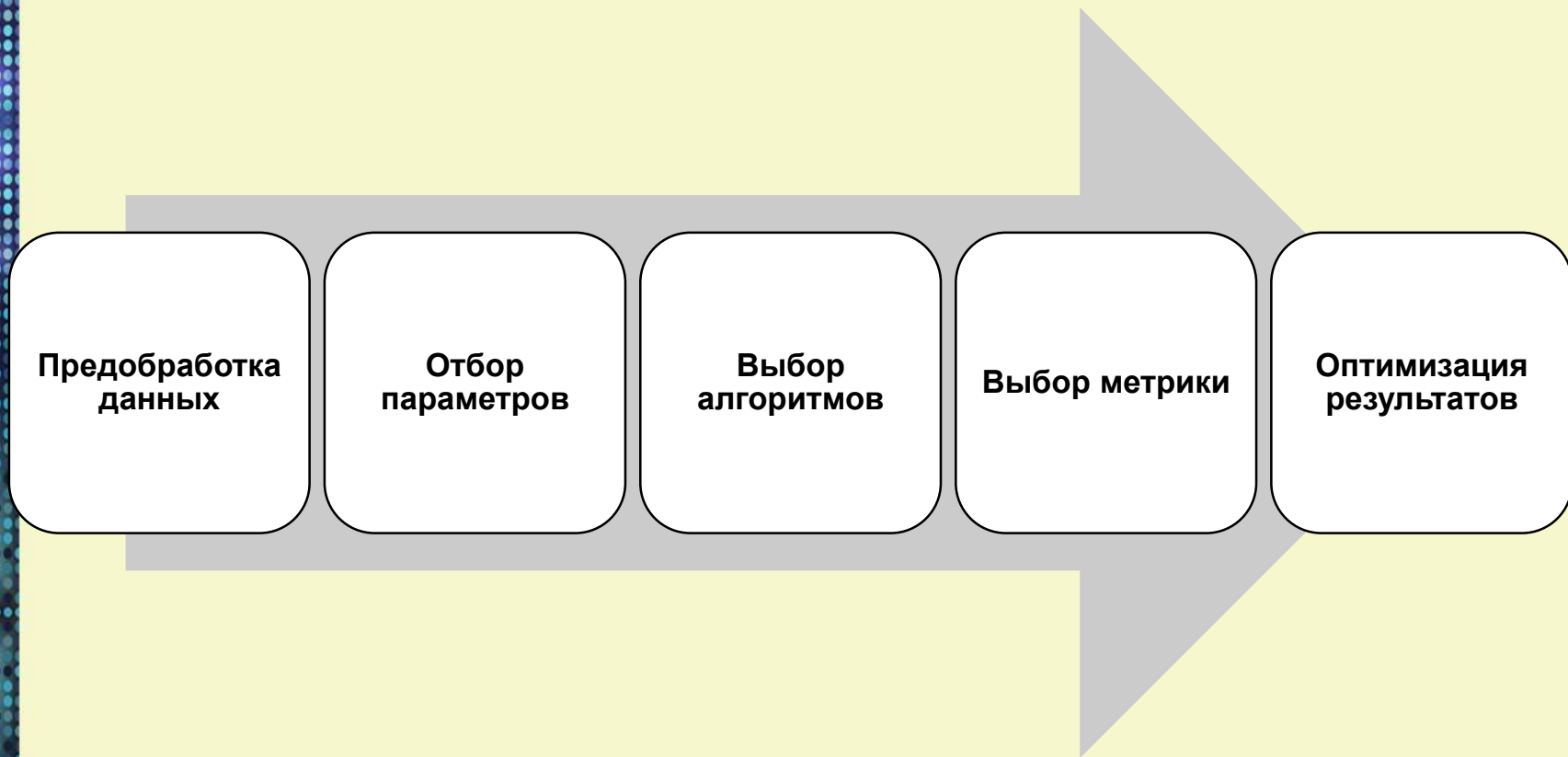


Задачи географии, картографии и геоинформатики, решаемые средствами машинным обучением

Особенности:

- результатом, как правило, должны быть несколько связанных величин;
- более сложная визуализация;
- интеграция с ГИС

Типовой порядок действий



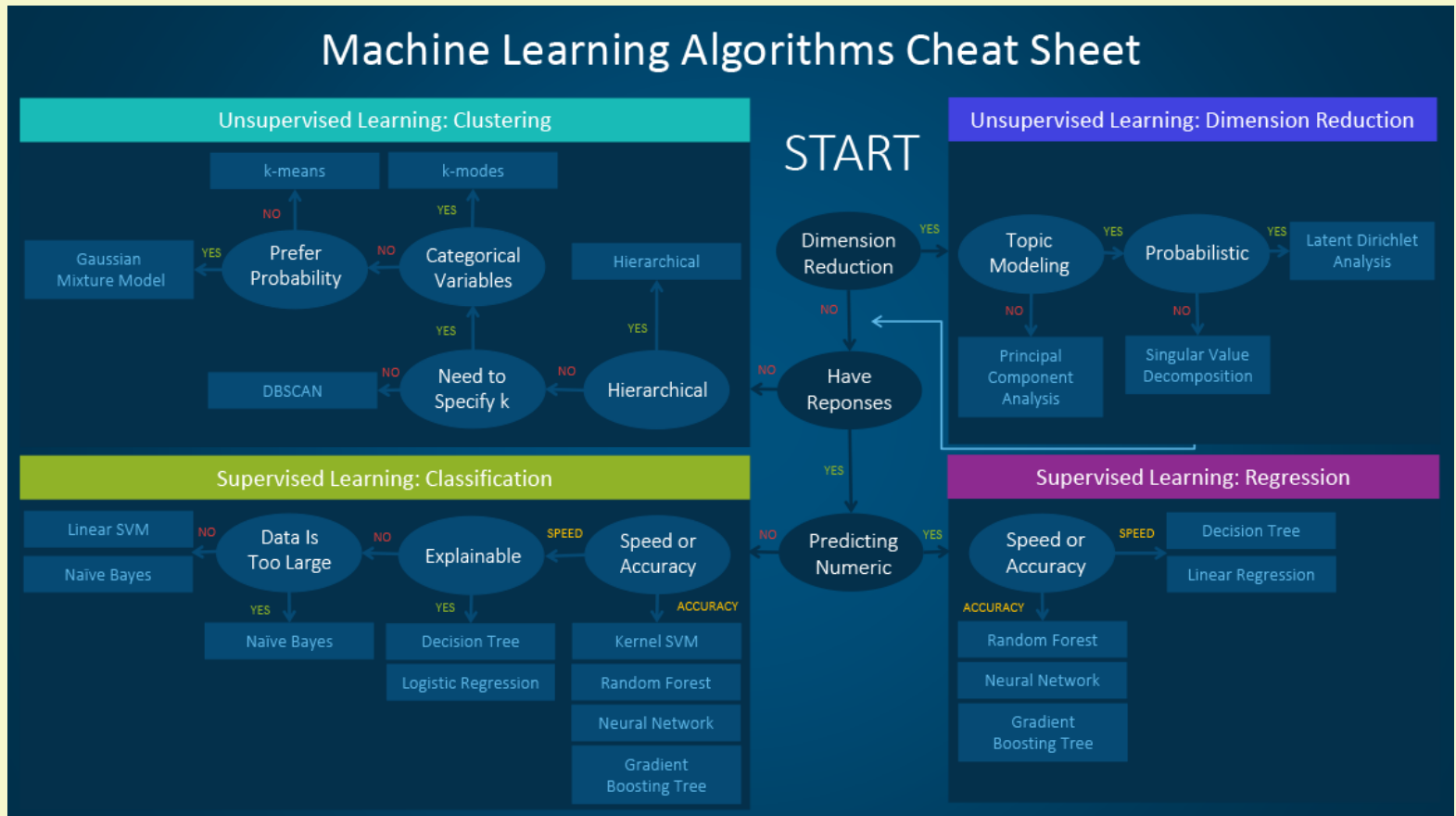
Предобработка пространственных данных

- ✓ проверка и назначение (при необходимости) правильных типов параметров (целое или дробное число, текст, дата и время и т.п.),
- ✓ заполнение пропущенных значений (если это возможно и в этом есть необходимость, поскольку некоторые алгоритмы не могут обрабатывать данные с пропусками),
- ✓ удаление ошибочных данных,
- ✓ различные типы кодирования
- ✓ ...

Выбор параметров

Способами отбора значимых параметров могут быть:

1. визуализация;
2. расчет значений корреляции (коэффициенты Пирсона, Спирмена и т.д.);
3. пространственная автокорреляция, хаотичность размещения объектов в пространстве (индекс Морана и т.д.);
4. расчет энтропии, в том числе для временных рядов (характерный показатель Ляпунова, коэффициент Хёрста, detrended fluctuation analysis и т.д.).



Which machine learning algorithm should I use? *By Hui Li, Principal Staff Scientist, Data Science, at SAS.*

<https://www.datasciencecentral.com/profiles/blogs/which-machine-learning-algorithm-should-i-use>

Выбор алгоритмов

- детекция и сегментация объектов на растровых данных;
 - деревья решений;
 - комбинация деревьев решений и градиентного бустинга (данный вариант является одним из наиболее универсальных решений, наряду с нейронными сетями);
 - традиционные и сверточные нейронные сети (являются наиболее универсальным решением, но требуется подбирать большое число параметров и подбирать архитектуру сети);
- классификация данных;
 - деревья решений и случайный лес;
 - нейронные сети;
 - метод опорных векторов;
 - логистическая классификация;
 - наивная байесовская классификация;

Выбор алгоритмов

- расчет пропущенных. либо прогнозных значений;
 - деревья решений и случайный лес;
 - комбинация деревьев решений и градиентного бустинга;
 - линейная регрессия;
 - логистическая регрессия;
 - нейронные сети;
 - ridge регрессия;
 - lasso регрессия;
 - метод опорных векторов;
 - кригинг и кокригинг (этот и следующие методы реализованы в большинстве геоинформационных систем);
 - метод обратных взвешенных расстояний;
 - методы локальных и глобальных полиномов;

Выбор алгоритмов

- прогнозирование и анализ временных рядов (для этих задач характерно наличие пространственных и атрибутивных данных для одной и той же территории в нескольких временных интервалах);
 - комбинация деревьев решений и градиентного бустинга
 - ARIMA/SARIMA;
 - нейронные сети глубокого обучения;
 - нейронные сети с долгой краткосрочной памятью (LSTM);
 - пространственно-временные повторяющиеся сверточные сети (SRCN);
 - пространственный анализ методом Монте-Карло (SMCA);
- поиск аномалий в данных;
 - метод опорных векторов;
 - метод главных компонент;
 - метод k-средних.

Метрики качества

Классификация:

Оценка модели:

F1-score

AUC ROC

...

Вероятность предсказания:

Mean average precision (MAP)

Logloss

...

Детекция/Сегментация:

Intersection over Union (IoU) / Jaccard coefficient (Коэффициент Жаккара)

Sorensen–Dice coefficient (коэффициент Сёренсена-Дайса) / F1-score,...

Пространственные данные:

коэффициент ранговой корреляции Kendall's Tau-b, процент попаданий в окружность заданного радиуса, коэффициент Джини,...

Способы улучшения качества

- ансамблирование или гибридные модели (boosting, bagging, stacking);
- перекрестная проверка (кросс-валидация);
- подбор параметров по сетке (grid search) и случайный подбор параметров (random search);

Внедрение

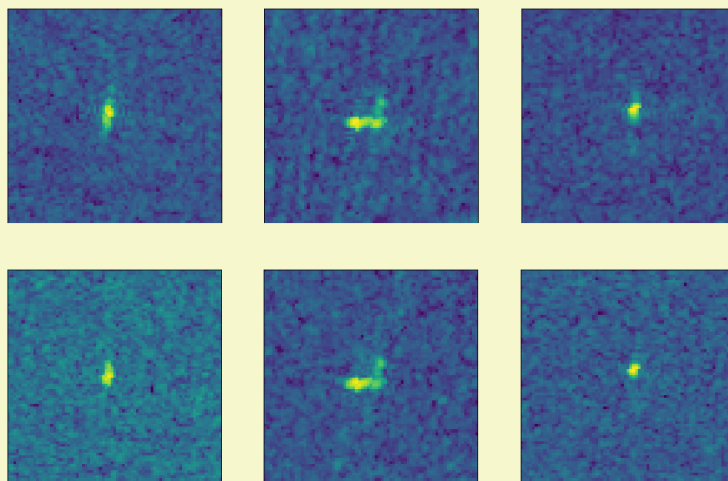
Варианты использования созданной модели на реальных данных:

- ✓ создание модулей для какой-либо ГИС;
- ✓ создание web-обвязки для выполнения скриптов на сервере;
- ✓ размещение на специализированном облачном сервисе.

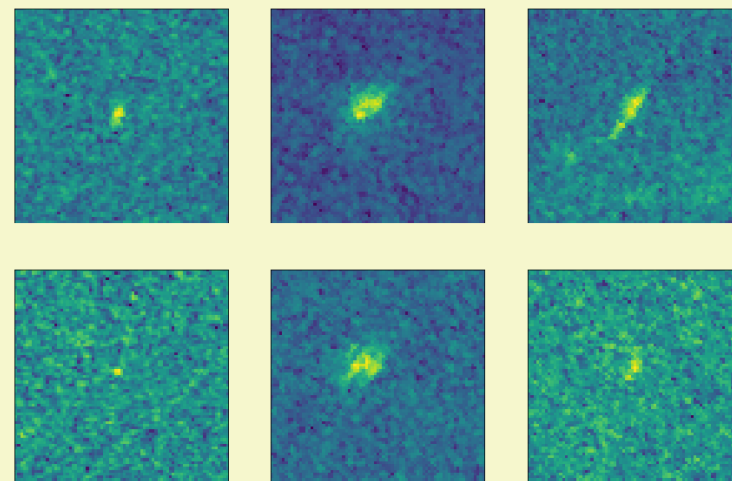
Iceberg Classifier Challenge

Снимки двух диапазонов
75x75 px – 5625 элементов
1604 – train set
8424 – test set
Метрика - logloss

Корабль



Айсберг



Iceberg Classifier Challenge

Логистическая регрессия	218s, 10,9s 0.6984, 1355 место
Градиентный бустинг XGBoost	186,5s, 14,2s 0.2682, 975 место
Ансамблирование градиентного бустинга XGBoost и LightGBM	2514,4s, 21,6s 0.2021, 536 место
Сверточная нейронная сеть Keras + Tensorflow	3800s, 39s 0.2497, 935 место
Предобученная сверточная нейронная сеть Keras + Tensorflow + VGG16	14175s, 48s 0.1745, 304 место
Предобученная сверточная нейронная сеть Keras + Tensorflow + InceptionV3	15867s, 53s 0.1780, 338 место

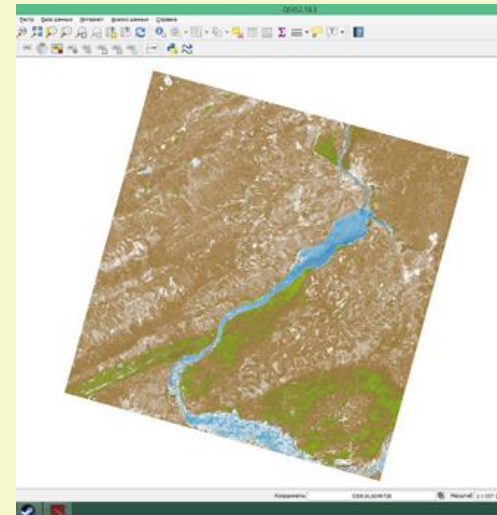
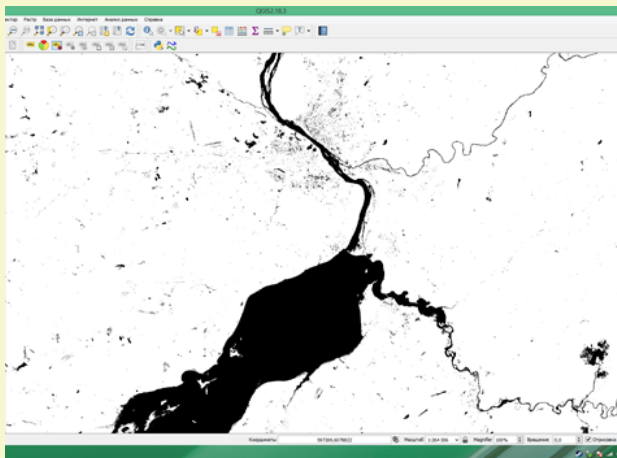
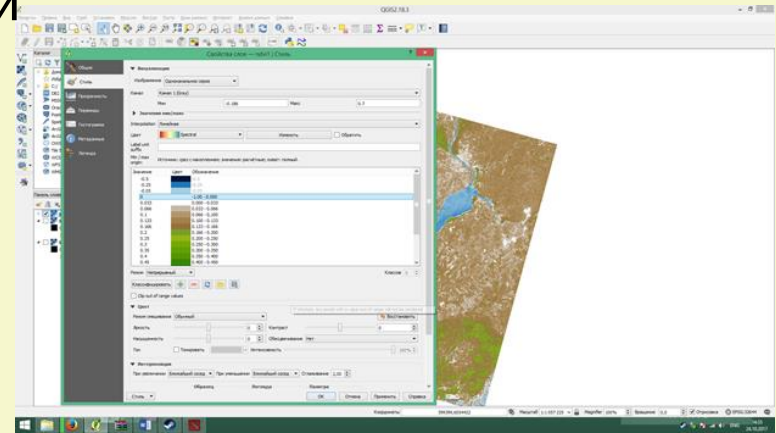
Сегментация объектов гидрографии

Участок Новосибирской области
площадью ~ 8000 км²

Landsat 8

QGIS

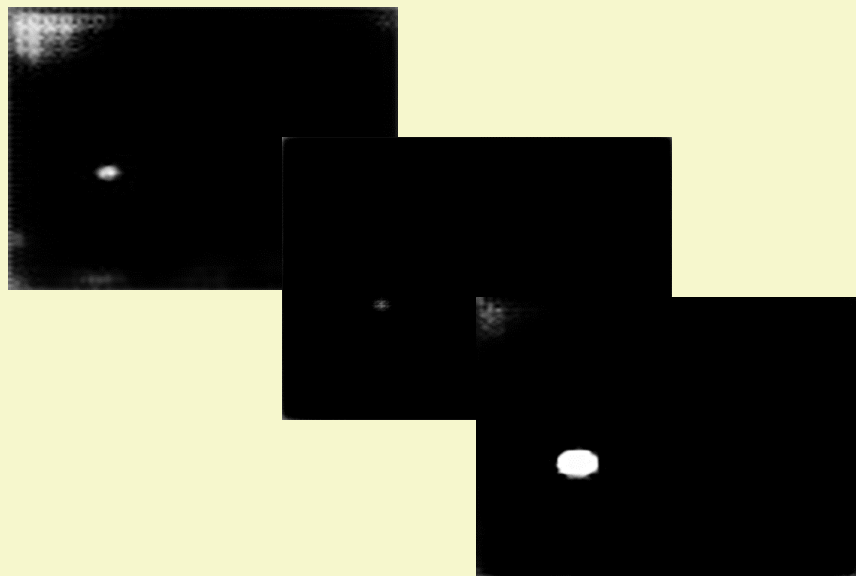
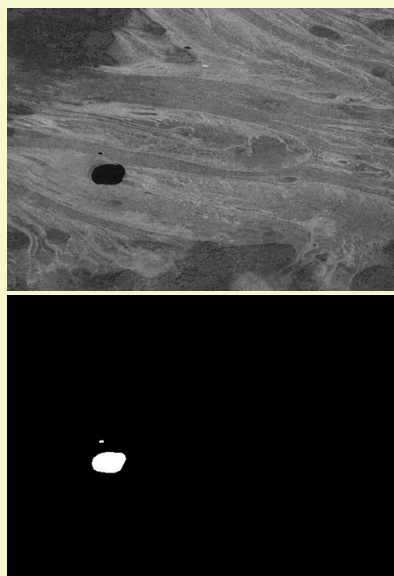
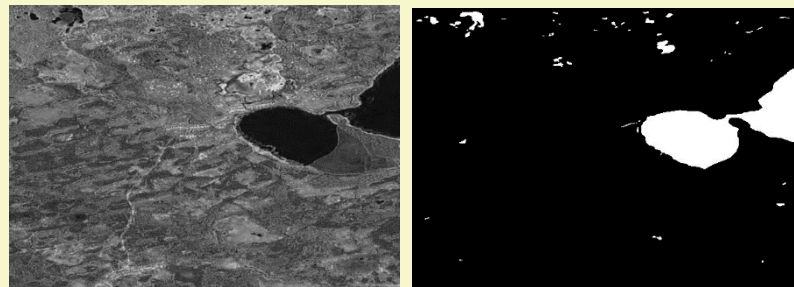
NDVI: Jaccard Index – 0,65



Сегментация объектов гидрографии

Keras + Tensorflow

Jaccard Index – от 0,40 до 0,75



DengAI: Predicting Disease Spread

Номер	Модель	Среднеквадратическая ошибка	Место в рейтинге
1	Линейная регрессия в Orange	29.8173	1024
2	KNN в Orange	33.8774	1348
3	Random Forest в Jupyter Lab	27.2981	876
4	Автоматический подбор параметров Random Forest в Jupyter Lab	26.5601	753
5	Random Forest в Orange	26.6130	758
6	XGboost в Jupyter Lab	27.8726	962
7	CatBoost в Jupyter Lab	37.1058	1812
8	LightGBM в Jupyter Lab	28.6947	963
9	Keras в Jupyter Lab после обработки данных инструментом StandardScaler, нейронная сеть без скрытых слоев	32.5481	1197
10	Keras в Jupyter Lab после обработки данных инструментом StandardScaler, нейронная сеть с двумя скрытыми слоями	27.3918	906
11	Автоматический подбор параметров Random Forest в Jupyter Lab с разделением по городам	26.3894	
12	Автоматический подбор параметров Random Forest в Jupyter Lab с разделением по городам без учета параметров климата и растительности	27.0361	
13	Автоматический подбор параметров Random Forest в Jupyter Lab с разделением по городам без учета параметров климата	27.2740	
14	Автоматический подбор параметров Random Forest в Jupyter Lab с разделением по городам без учета параметров растительности	27.5745	



Выводы

✓ машинное обучение работает

Выводы

✓ в любом случае будет результат

Выводы

✓ но он может быть плохим;

Выводы

- ✓ но, в ряде случаев, сильно лучше традиционного подхода и далее круг успешно решаемых задач будет увеличиваться;

Выводы

- ✓ решение пространственных задач средствами ML специфично и порог вхождения высок, но перспективы очень велики.



ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ ПРИ РЕШЕНИИ ГЕОИНФОРМАЦИОННЫХ ЗАДАЧ

Колесников А.А., доцент кафедры картографии и геоинформатики СГУГиТ
alexeykw@yandex.ru

«ИнтерКарто/ИнтерГИС», г. Петрозаводск, 2018