

# Индивидуальный подбор весов в ансамбле алгоритмов автоматического машинного обучения

---

Малышева Александра

Научный руководитель: Алексей Шпильман

# Актуальность

name	#	категориальные	численные
AdaBoost (AB)	4	1 (-)	3 (-)
Bernoulli naïve Bayes	2	1 (-)	1 (-)
decision tree (DT)	4	1 (-)	3 (-)
extreml. rand. trees	5	2 (-)	3 (-)
Gaussian naïve Bayes	-	-	-
gradient boosting (GB)	6	-	6 (-)
kNN	3	2 (-)	1 (-)
LDA	4	1 (-)	3 (1)
linear SVM	4	2 (-)	2 (-)
kernel SVM	7	2 (-)	5 (2)
multinomial naïve Bayes	2	1 (-)	1 (-)
passive aggressive	3	1 (-)	2 (-)
QDA	2	-	2 (-)
random forest (RF)	5	2 (-)	3 (-)
Linear Class. (SGD)	10	4 (-)	6 (3)

Классификаторы

name	#	категориальные	численные
extreml. rand. trees prepr.	5	2 (-)	3 (-)
fast ICA	4	3 (-)	1 (1)
feature agglomeration	4	3 (0)	1 (-)
kernel PCA	5	1 (-)	4 (3)
rand. kitchen sinks	2	-	2 (-)
linear SVM prepr.	3	1 (-)	2 (-)
no preprocessing	-	-	-
nystroem sampler	5	1 (-)	4 (3)
PCA	2	1 (-)	1 (-)
polynomial	3	2 (-)	1 (-)
random trees embed.	4	-	4 (-)
select percentile	2	1 (-)	1 (-)
select rates	3	2 (-)	1 (-)
one-hot encoding	2	1 (-)	1 (1)
imputation	1	1 (-)	-
balancing	1	1 (-)	-
rescaling	1	1 (-)	-

Препроцессинг

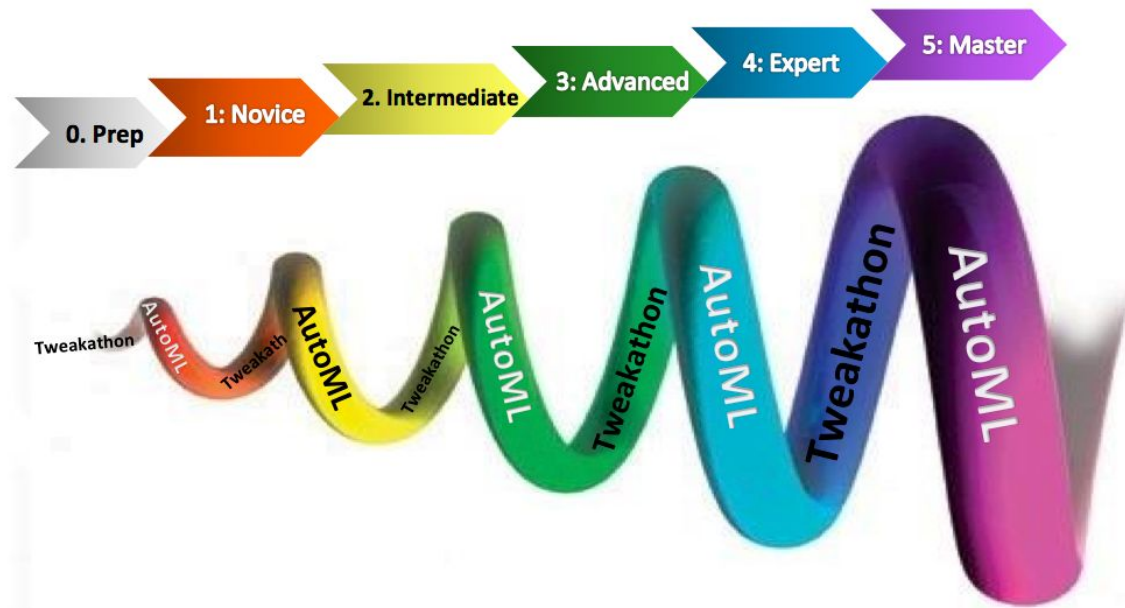
# Существующие библиотеки

Spark Machine learning library

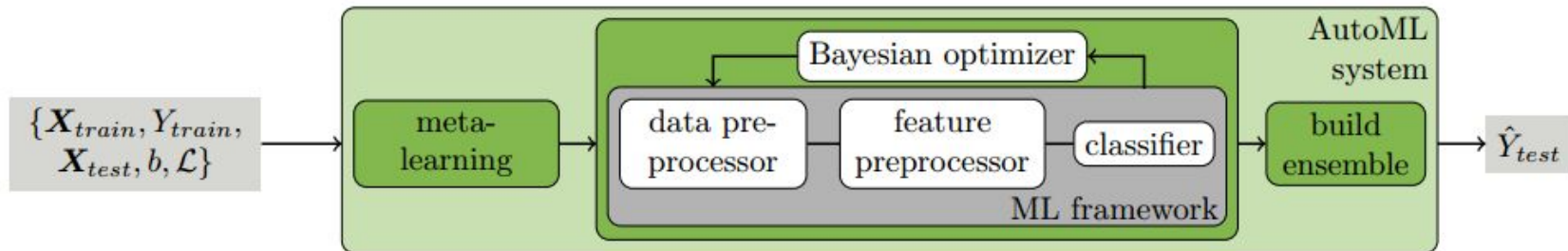
**Auto-sklearn**

Auto-WEKA

Auto-Net



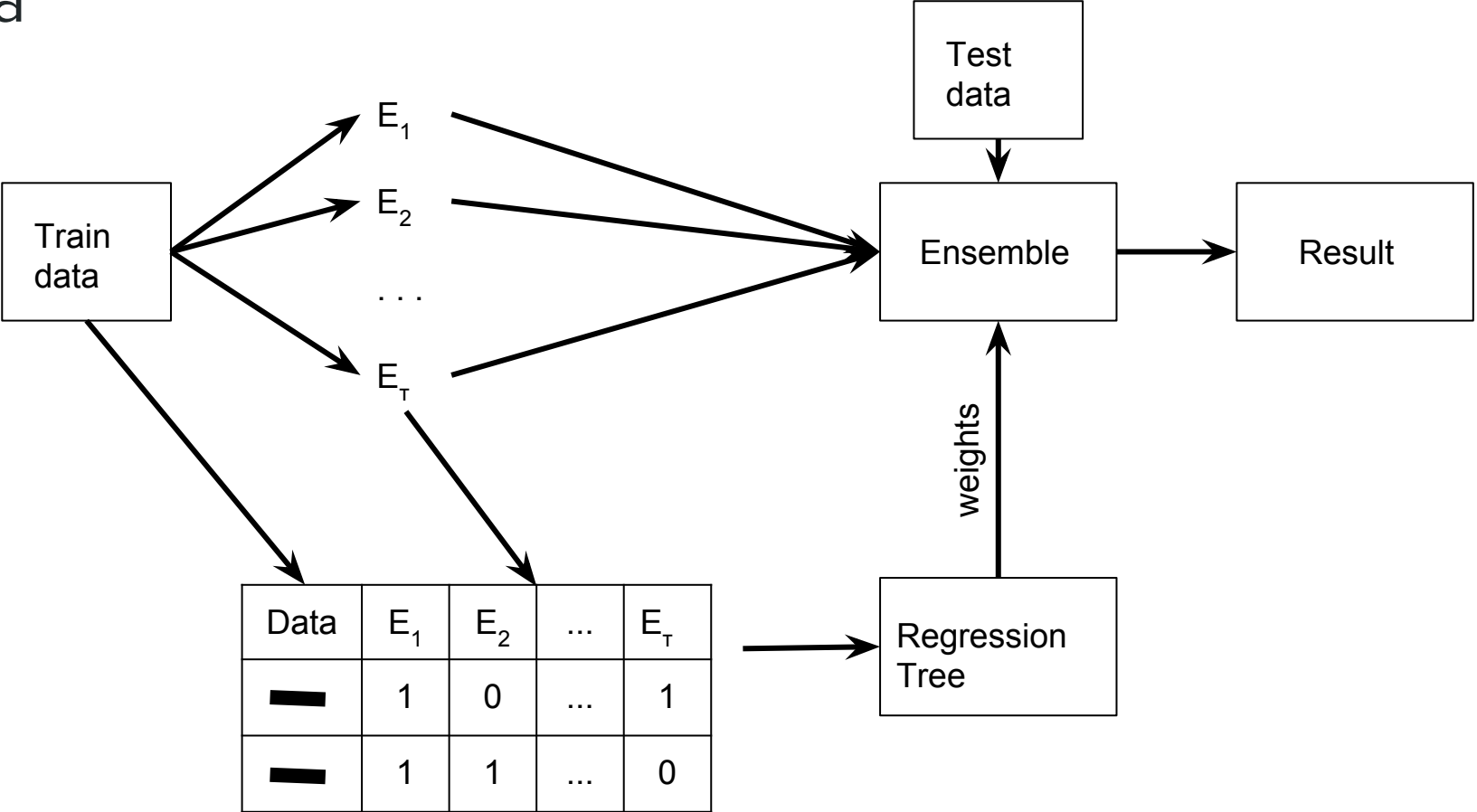
# Auto-sklearn



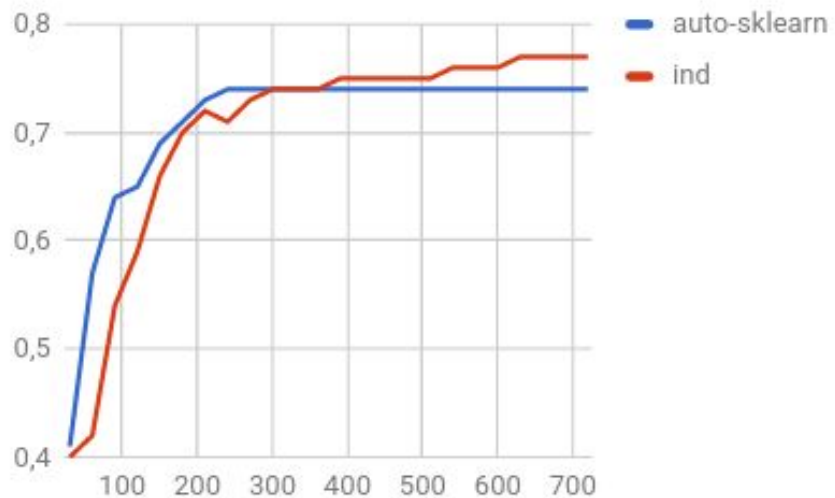
# Постановка задачи

- Реализовать алгоритм индивидуального подбора весов в ансамбле алгоритмов автоматического машинного обучения
- Провести сравнение с уже имеющейся библиотекой автоматического машинного обучения на разных данных

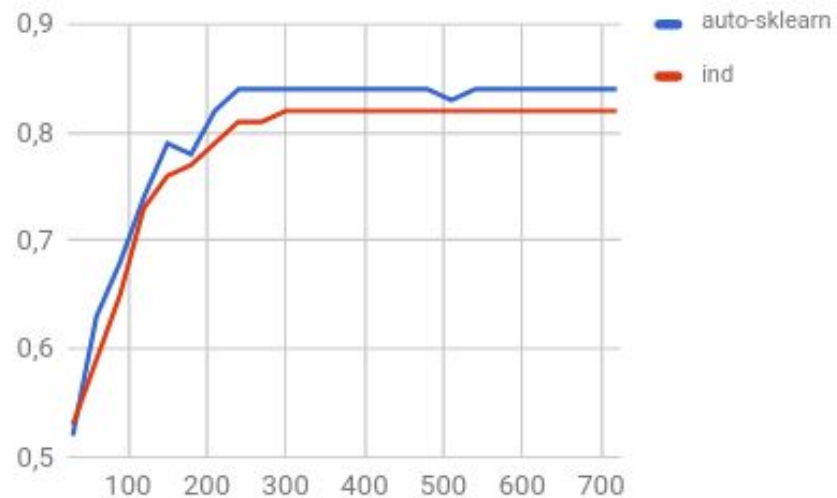
# Схема



# Результаты

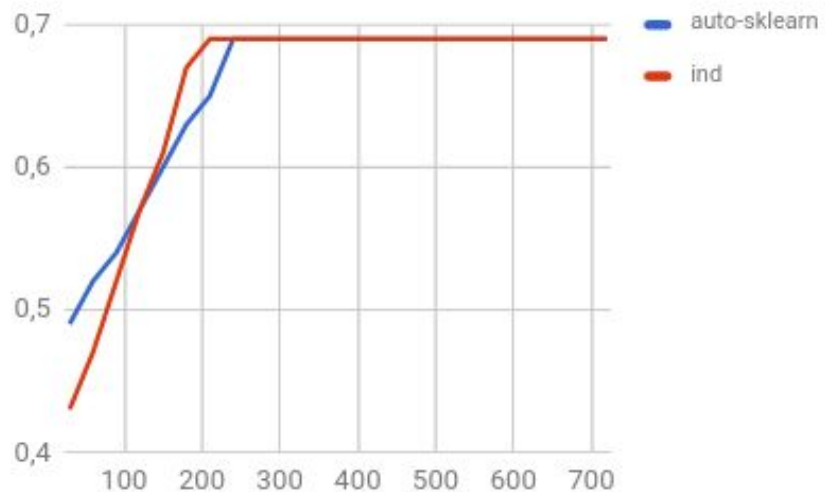


Ionosphere Data Set

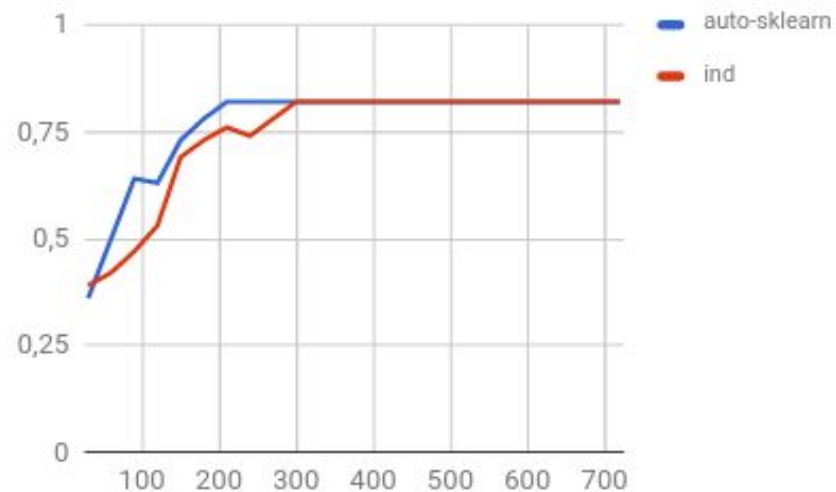


Contraceptive Method Choice Data Set

# Результаты



Breast Cancer Data Set



Hepatitis Data Set



# Результаты

	# classes	# attributes	# samples	auto-sklearn 120	ind 120	auto-sklearn	ind
Ionosphere	2	34	351	<b>0,65</b>	0,59	0,74	<b>0,77</b>
Cylinder Bands	2	39	512	<b>0,69</b>	0,57	0,87	<b>0,89</b>
Lung Cancer	4	56	32	<b>0,82</b>	0,73	0,82	0,82
Hayes-Roth	3	5	160	<b>0,79</b>	0,58	0,81	0,81
Hepatitis	9	19	155	<b>0,63</b>	0,53	0,82	0,82
Breast Cancer	2	24	286	0,57	<b>0,58</b>	0,69	0,69
Flags	8	30	194	0,48	0,48	0,89	0,89
Glass Identification	7	10	214	0,67	<b>0,68</b>	0,81	0,81
Dermatology	6	38	366	<b>0,56</b>	0,53	0,64	0,64
Credit Approval	2	15	690	<b>0,64</b>	0,46	<b>0,87</b>	0,86
Contraceptive Method Choice	5	19	1473	<b>0,74</b>	0,73	<b>0,84</b>	0,82
Heart Disease	2	19	303	<b>0,64</b>	0,48	<b>0,89</b>	0,87

# Планы

- Обработка численных предсказаний классификатора
- Векторизация алгоритмов
- Оптимизация алгоритмов