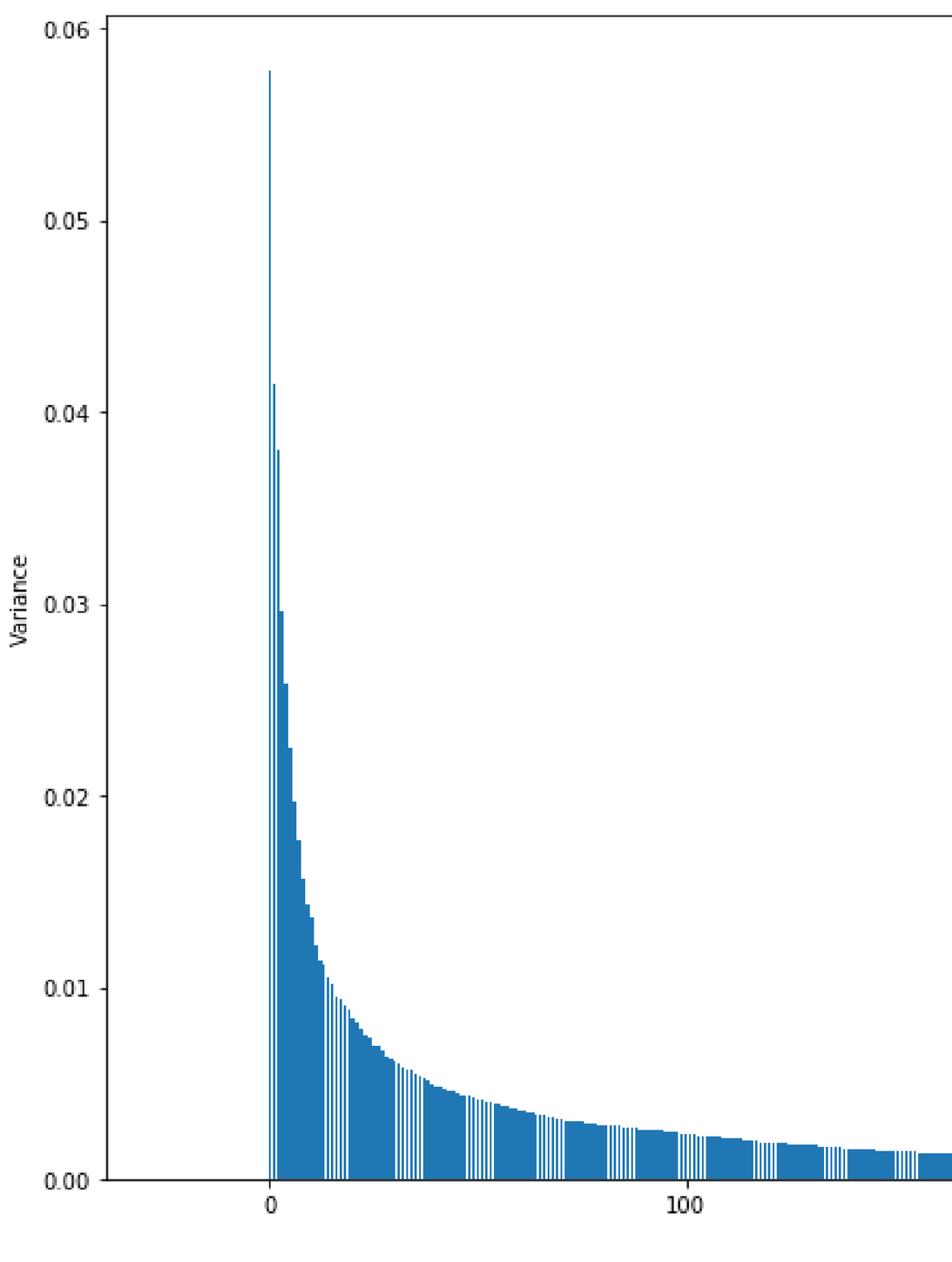
Задача классификации

Accuracy: 0.9744

Место: 1514 из 2271

Логин: iKintosh

АЛЕКСЕЕВ АНДРЕЙ



DIGIT RECOGNIZER

ДАТАСЕТ

Digit Recognizer

Learn computer vision fundamentals with the famous MNIST data

k kaggle

ЦЕЛЬ: ПРЕДСКАЗАТЬ ВЕРНУЮ МЕТКУ КЛАССА

ИТОГОВОЕ РЕШЕНИЕ

Обработка данных	Модель
Сжатие признакового пространства при помощи PCA до 50 компонент	SVM с RBF kernel Параметры: gamma: 1 / (n_features * X.var()) C: 10

Работа с признаками

Подход	Качество на **CV**
ДАТАСЕТ БЕЗ ИЗМЕНЕНИЙ	**SVM: 0.9515** LogReg: 0.8692 RandomForest: 0.9415 KNN: 0.9344
ЗАМЕНА ЗНАЧЕНИЙ >0 НА 1	**SVM: 0.9488** LogReg: 0.8942 RandomForest: 0.9465
PCA: СОХРАНЕНИЕ 90% VARIANCE	**SVM: 0.9498** LogReg: 0.8996 KNN: 0.9069
PCA: СОХРАНЕНИЕ 75% VARIANCE	**SVM: 0.9471** LogReg: 0.8909 KNN: 0.9226
PCA: СЖАТИЕ ДО 50 ПРИЗНАКОВ	**SVM: 0.9467** LogReg: 0.8926 RandomForest: 0.9145 KNN: 0.9232

*Обучение при проверке различных видов датасета велось на подвыборке в 5000 элементов (это было сделано для ускорения скорости работы)

Опробованные методы

Метод	LogReg	AdaBoost	KNN	SVM	RandomForest
Настраиваемые параметры	**C (сила регуляризации)**	**N алгоритмов**	**N соседей, метрика расстояния, размер листьев**	**Ядро, Гамма**	**N деревьев, Глубина**
Область настройки	**0.1-2**	**1-200**	**1-200** ***1, 2*** ***5-100***	**poly, rbf** ***scale, auto***	**1-500** ***2-None***
Результат на кросс-валидации (выбран лучший)	**0.8996**	**0.7022**	**0.9344**	**<u>0.9515</u>**	**0.9465**
Результат на лидерборде (был отправлен только лучший алгоритм	*-//-*	*-//-*	*-//-*	**0.9744** ***(1514 из 2271)***	*-//-*

*Также был опробован подход с SoftVoting, но он работал не лучше, чем SVM