

УДК 004.9, 004.5, 004.41/.42, 004.43  
МРНТИ 50.41.25

## СРАВНЕНИЕ АЛГОРИТМОВ ОБУЧЕНИЯ С УЧИТЕЛЕМ И БЕЗ УЧИТЕЛЯ ДЛЯ ОБНАРУЖЕНИЯ ДЕПРЕССИВНЫХ ПОСТОВ

С.С. НАРЫНОВ, Д. МУХТАРХАНУЛЫ, И.М. КЕРИМОВ

ТОО "AlemResearch"

**Аннотация:** Согласно последним данным ВОЗ, опубликованным в 2017 году, количество самоубийств в Казахстане составило 4855 или 3,55% от общего числа смертей. Уровень смертности с поправкой на возраст составляет 27,74 на 100 000 населения. Казахстан занимает четвертое место в мире по этому показателю. В этой статье представлено сравнение алгоритмов машинного обучения с учителем и без учителя для выявления депрессивного контента в постах в социальных сетях с акцентом на безнадежность и психологическую боль для семантического анализа в качестве ключевых причин самоубийства. Самоубийство не является спонтанным, и подготовка к самоубийству может длиться около года, в течение которого человек будет демонстрировать признаки своего состояния, в нашем случае, публикуя депрессивный контент в своем профиле в социальной сети. Этот алгоритм помогает в обнаружении депрессивного контента, который может вызвать самоубийство, чтобы помочь людям найти уверенную помощь от психологов национального центра по предотвращению самоубийств в Казахстане. Получив наивысший результат для 95% оценки  $f1$  для случайного леса (обучение с учителем) с моделью векторизации  $tf-idf$ . В заключение можно сказать, что алгоритм  $K$ -means (обучение без учителя) с использованием  $tf-idf$  показывает впечатляющие результаты, которые ниже только на 4%  $f1$  и точности.

**Ключевые слова:** случайный лес, анализ тональности,  $k$ -средние, машинное обучение, обучение с учителем, обучение без учителя

## COMPARISON OF SUPERVISED LEARNING WITH UNSUPERVISED LEARNING ALGORITHMS IN DEPRESSIVE POST DETECTION

**Abstract:** According to the latest WHO data published in 2017 Suicide Deaths in Kazakhstan reached 4,855 or 3.55% of total deaths. The age adjusted Death Rate is 27.74 per 100,000 of population ranks Kazakhstan #4 in the world. This article shows the comparison of supervised and unsupervised machine learning algorithms, for detecting of depressive content in posts in social networks with emphasis on hopelessness and psych-ache for semantic analysis as the key reasons for suicide. Suicide is not an impulsive act and preparation for suicide can last about a year, during which a person will show signs of his condition in our case posting depressive content on his social network profile. This algorithm helps in detections of depressive content which can cause suicide, to help founded persons reach confident help from psychologists of national suicide preventing center in Kazakhstan. Obtaining highest result for 95% of  $f1$ -score for Random Forest(supervised) with  $tf-idf$  vectorization model, in conclusion of comparison we may say that  $K$ -means(Unsupervised) using  $tf-idf$  shows impressive results, which is only 4% lower in  $f1$ -score and precision.

**Keywords:** random forest, sentiment analysis,  $k$ -means, machine learning, supervised learning, unsupervised learning

## БАҚЫЛАНАТЫН ЖӘНЕ БАҚЫЛАНБАЙТЫН ОҚЫТУДЫ ДЕПРЕССИЯЛЫҚ ПОСТТАРДЫ АНЫҚТАУ ҮШІН АЛГОРИТМДЕРДІ САЛЫСТЫРУ

**Аңдатпа:** 2017 жылы жарияланған Дүниежүзілік денсаулық сақтау ұйымының соңғы деректеріне сәйкес, Қазақстанда өз-өзіне қол жұмсау өлімінің жалпы саны 4 855-ге жетті немесе жалпы өлімнің 3,55% -ын құрайды. Өлімнің деңгейі 100,000 адамға шаққанда 27,74 болса, Қазақстан әлемдегі № 4 орынға шықты. Бұл мақалада әлеуметтік және әлеуметтік желілердегі депрессиялық контентті анықтау үшін, белгіленген және белгіленбеген машина алгоритмдерін салыстыру ұсынылды. Бұл семантикалық талдауға психологиялық көмек көрсету, өз-өзіне қол жұмсаудың негізгі себептері болып табылады. Өзін-өзіне қол жұмсау – бұл импульсивтік әрекет емес, ал өзіне-өзі қол жұмсау үшін шамамен дайындық бір жылға созылады, оның барысында адам біздің әлеуметтік желі профилінде депрессиялық контентті орналастыру жағдайында оның жан-күйзелісін көрсететін белгілері болады. Бұл алгоритм, өзіне-өзі қол жұмсауды тудыруы мүмкін депрессивтік контентті анықтауға көмектеседі, негізін қалаушыларға Қазақстандағы ұлттық суицидтік алдын алу орталығының психологтарынан сенімді көмекке қол жеткізуге жәрдемдеседі. Tf-idf векторизациясының үлгісімен бақыланатын RandomForest үшін f1-score 95% жоғары нәтиже алу, салыстыруды аяқтағаннан кейін, tf-idf-ты пайдаланудың K-means белгіленген алгоритмі әсерлі нәтижелерді көрсетеді, бұл тек 4% төмен f1-score және precision.

**Түйінді сөздер:** random forest, реңкілікті анықтау, k-means, machine learning, бақыланатын оқыту, бақыланбайтын оқыту

### ВВЕДЕНИЕ

Ежегодно во всем мире от самоубийств умирает более миллиона человек. Глобальный уровень самоубийств составляет 16 на 100 000 населения [1]. В среднем, один человек умирает от самоубийства каждые 40 секунд где-то в мире. Глобальные показатели самоубийств выросли на 60% за последние 45 лет. Согласно последним данным ВОЗ, опубликованным в 2017 году, смертность от самоубийств в Казахстане достигла 4855 или 3,55% от общего числа смертей [2]. Уровень смертности с поправкой на возраст составляет 27,74 на 100 000 населения. Казахстан занимает 4 место в мире.

Это социальное заболевание является результатом различных причин, однако в этом исследовании мы сосредоточились на основных причинах самоубийства, согласно статье Клонского Е.Д. [3], это безнадежность и психическая боль, эти ключевые ценности получили наибольшее место в анализе Из депрессивных публикаций людей в социальных сетях с доказанной депрессией в диагностике психологических экспертов наша предыдущая работа была связана с обнаружением суицидальных постов [4].

Мы собрали 35000 сообщений на русском языке для людей с диагностированной депрессией разного уровня: тяжелое, хроническое, маниакальное, постоянное депрессивное расстройство и так далее. Также мы использовали около 50000 личных постов из социальных сетей с негативными настроениями на различные темы для создания однородной текстовой базы, разделенной на тренировочную базу для обучения алгоритмов и тестовые базы для оценки производительности алгоритмов.



Рис. 1 – Разделение данных

Теперь мы сравниваем производительность двух основных типов алгоритмов машинного обучения в рамках одной области идентификации депрессивных постов с областью для безнадежности и психической боли, выраженной в тексте. В результате экспериментов были получены f1-score 95% и ROC-площадь 0,98 со случайным лесом с моделью векторизации tf-idf, при этом контролируются параметры и перекрестная проверка алгоритмов с помощью кривой ROC. Сравнивая с нашим ранее построенным алгоритмом, мы увеличили прогноз почти на 20% [4].

Сравнение контролируемых и неконтролируемых алгоритмов обучения показало, что неконтролируемые алгоритмы могут приводить к очень высоким f1-score с точностью до 90% при обнаружении депрессивного контента в постах пользователей социальных сетей.

#### ПОХОЖИЕ РАБОТЫ

*Прогнозирование суицидальных тенденций в данных Twitter с использованием алгоритмов машинного обучения (Marouane Birjali 2016)*

Эта конкретная работа, связанная с Marouane Birjali, рассказывает в этой статье о предложении системы обнаружения суицидальных мыслей, предсказывающей суицидальные действия с использованием данных Twitter, которые могут автоматически анализировать настроения этих твитов [5]. Они исследуют инструмент извлечения данных для извлечения полезной информации для классификации твитов, собранных из Twitter на основе алгоритмов классификации машинного обучения. И приведите экспериментальные результаты, показывающие, что их метод обнаружения суицидальных актов с использованием данных Twitter и алгоритмы машинного обучения проверяют эффективность работы с точки зрения отзыва, точности и точности анализа настроений.

Основное отличие нашей работы связано с объемом анализа, в то время как наши коллеги использовали данные Twitter, которые состоят из коротких 140-символьных сообщений в виде заголовков, мы решили использовать VK.com, Facebook.com и uvision.kz. социальные платформы, где пользователи

не ограничены в объеме текста, который они могут публиковать на своих страницах, таким образом, анализ текста, показывающий нам более определенные результаты, поэтому мы собрали из этих источников 35000 сообщений с доказанной депрессией авторов этих сообщений, этот набор данных используется для обучения и тестирования различных алгоритмов ИИ с контролируемым и неконтролируемым обучением.

*Раннее выявление самоубийств с использованием аналитики больших данных в режиме реального времени (Хардик А. Пател, 2016 г.)*

В этой статье представлено приложение, в котором предиктивный анализ используется для сбора комментариев и постов в социальных сетях для выявления лиц на этих сайтах, подверженных суицидальным мыслям и тенденциям. Представлена также модель, в которой для анализа идентификаторов больших данных и сентиментальной аналитики использовалась ветвь. профили пользователей с сообщениями о самоубийстве.

Они пришли к выводу, что изобретение – это инструмент для агентств по предотвращению самоубийств, которые могут использовать его для прогнозного анализа с использованием больших данных, выявления лиц, страдающих суицидальными тенденциями, что помогает им своевременно вмешиваться и спасти жизни.

В отличие от работы Хардика Пателя, мы не пытаемся предложить систему, но мы постоянно совершенствуем наши алгоритмы, сравнивая их друг с другом, используя решения и всемирно известные алгоритмы ИИ [6].

*Изучение различных стратегий разметки сообщений о самоубийствах в социальных сетях: экспериментальное исследование (Liu, et al. 2017)*

В статье Лю и Тонга делается попытка дать ответ о том, как получить надежные размеченные данные, которые предполагают авторы, что, по их мнению, во многом зависит от того, что спрашивают аннотаторы, и какую часть данных они размечают [7]. Они провели несколько раундов разметки дан-

ных и собрали аннотации от краудсорсинговых работников и работают с основными экспертами. Полученные ярлыки были объединены различными способами для обучения серии алгоритмов на основе обучения с учителем.

Их предварительные оценки показывают, что использование единодушно согласованных меток от нескольких аннотаторов полезно для создания надежных моделей машин. Хотя эта статья, кажется, является результатом серьезного исследования, такого же, как и в предыдущей статье HardikPatel, это исследование основано на постах в Твиттере также в качестве постов на английском языке, в нашей работе мы используем VK, FB, Twitter и различные популярные Казахстан социальные сети устраняют недостаток текстовой базы, что, по нашему мнению, может привести к снижению качества определения депрессивного компонента в выражениях [6].

## МЕТОДОЛОГИЯ

### *Анализ и сбор данных*

Поскольку данные о лицах, совершивших самоубийство, являются конфиденциальными, мы собирали депрессивные сообщения пользователей социальных сетей. Была проделана огромная работа по созданию парсера, который собирает посты от публичных групп с депрессивным или суицидальным содержанием. Мы собрали около 35 000 депрессивных сообщений и 50 000 обычных сообщений, таких как новости, блоги, добавления и т.д.

### *Анализ депрессивных постов*

Основываясь на тщательном анализе текстового корпуса и основных тем и эмоций, наши психологи выделили три возможных причины самоубийства:

1. Необходимость поддержки – отчаянная попытка привлечь внимание других людей к их психическому состоянию.
2. Избегание - неспособность терпеть любые дальнейшие невыносимые душевные боли, вину или стыд за социально неприемлемые действия.
3. Протест - это протест против сложных семейных проблем, часто связанных с выражением эмоций гнева и обвинения. Письмен-

ная записка часто адресована конкретному человеку или группе людей.

Самоубийство не является импульсивным актом, и подготовка к самоубийству может длиться около года, в течение которого человек будет демонстрировать признаки своего состояния. Мы используем анализ сентимента, чтобы обнаружить этот опасный период.

## ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА И АНАЛИЗ

Для обнаружения депрессивных публикаций с контролируемым обучением мы использовали RandomForest, GradientBoosting [4]. В то время как для обучения без учителя мы использовали K-средние с 2-х кластерной моделью.

### *Предобработка данных*

Во-первых, все тексты были лемматизированы – процесс удаления только окончаний и возврата базовой или словарной формы слова, которая известна как лемма. Для лемматизации слов в контексте русского языка использовался лемматизатор «MyStem» от Яндекса, поскольку он продемонстрировал отличные результаты. Впоследствии, nltk-библиотека для стоп-слов использовалась для удаления стоп-слова, следовательно, снижение потенциального шума в данных. Также были удалены цифры, специальные символы, не кириллические буквы.

Во-вторых, предварительно обработанные тексты были векторизованы - процесс представления текстов в векторном пространстве для арифметических операций над всей структурой данных. Векторный вид экономит время. Для векторизации текстов модели TF-IDF и Word2Vec были использованы.

TF-IDF расширяется как TermFrequency-InverseDocumentFrequency, который в основном говорит о важности слова в корпусе или наборе данных. TF-IDF содержат две концепции: термин-частота (TF) и обратная частота документа (IDF)

Word2vec - это техника глубокого обучения с двухслойной нейронной сетью. Google-Word2vec берет данные из больших данных и преобразует их в векторное пространство.

Word2vec в основном помещает слово в пространство признаков таким образом, что их местоположение определяется их значением, то есть слова, имеющие сходное значение, группируются вместе, и расстояние между двумя словами также имеет одинаковое значение.

**Обучение с учителем**

Для эксперимента были протестированы 2 лучших алгоритма согласно нашей ранее опубликованной работе, а также с векторизацией tf-idf [4]:

1. Градиентный бустинг с помощью word2vec
2. Случайный лес с word2vec
3. Градиентный бустинг с помощью tf-idf
4. Случайный лес с tf-idf

**Градиентный бустинг с помощью word2vec**

Accuracy 0.9056777117657867				
Metrics	precision	recall	f1-score	support
0.0	0.95	0.92	0.93	11811
1.0	0.88	0.87	0.84	4516
avg / total	0.91	0.91	0.91	16327

Precision: 91%  
 Recall: 91 %  
 F1-score:91%

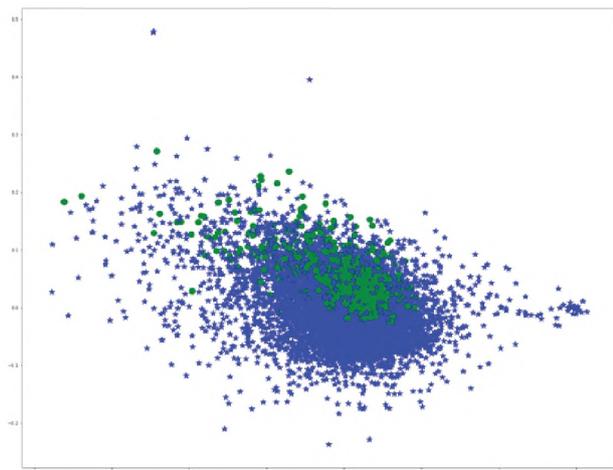


Рис. 2 – Графическое представление векторов word2vec в 2D-пространстве, где зеленые метки - это депрессивные посты, а синие метки – обычные посты

**Случайный лес с word2vec**

Accuracy 0.895081766399216				
Metrics	precision	recall	f1-score	support
0	0.96	0.90	0.93	12212
1	0.75	0.89	0.81	4115
avg / total	0.91	0.90	0.90	16327

Precision: 91%  
 Recall: 90 %  
 F1-score:90%

**Градиентный бустинг с помощью tf-idf**

Accuracy 0.954125068904269				
Metrics	precision	recall	f1-score	support
0.0	0.99	0.95	0.97	11850
1.0	0.88	0.96	0.92	4477
avg / total	0.96	0.95	0.95	16327

Precision: 96%  
 Recall: 95 %  
 F1-score:95%

**Случайный лес с tf-idf**

Accuracy 0.9553500336865315				
Metrics	precision	recall	f1-score	support
0	0.98	0.96	0.97	11658
1	0.90	0.95	0.92	4669
avg / total	0.96	0.96	0.96	16327

Precision: 96%  
 Recall: 96 %  
 F1-score:96%

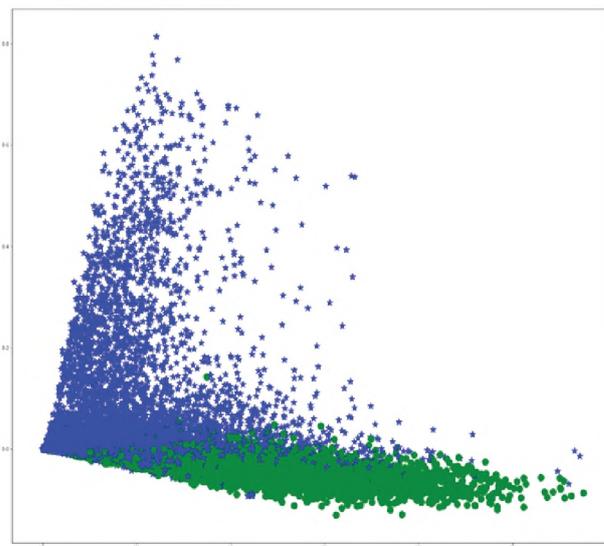


Рис. 3 – Графическое представление векторов tf-idf в 2D-пространстве, где зеленые метки - это депрессивные посты, а синие метки – обычные посты

**Обучение без учителя**

Для эксперимента были протестированы

К-средние с tf-idf и word2vec:

1. К-средние с word2vec
2. К-средние с tf-idf

**К-средние с word2vec**

Metrics on training set				
	precision	recall	f1-score	support
0	0.73	0.85	0.78	39268
1	0.69	0.52	0.60	26037
avg / total	0.71	0.72	0.71	65305

Metrics on testing set				
	precision	recall	f1-score	support
0	0.73	0.85	0.79	9801
1	0.71	0.53	0.60	6526
avg / total	0.72	0.72	0.71	16327

Precision: 72%

Recall: 72%

F1-score: 71%

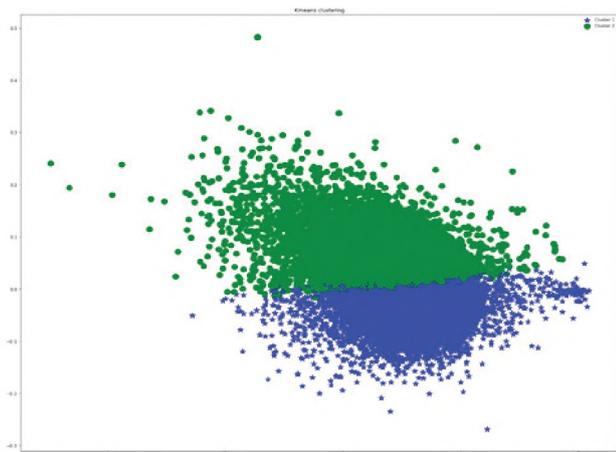


Рис. 4 – Графическое представление векторов word2vec в 2D-пространстве, где зеленые метки - это депрессивные посты, а синие метки - обычные посты

**К-средние с tf-idf**

Metrics on training set				
	precision	recall	f1-score	support
0	0.95	0.91	0.93	47763
1	0.79	0.88	0.83	17542
avg / total	0.91	0.90	0.91	65305

Metrics on testing set				
	precision	recall	f1-score	support
0	0.96	0.91	0.93	11975
1	0.79	0.88	0.83	4352
avg / total	0.91	0.91	0.91	16327

Precision: 91%

Recall: 91%

F1-score: 91%

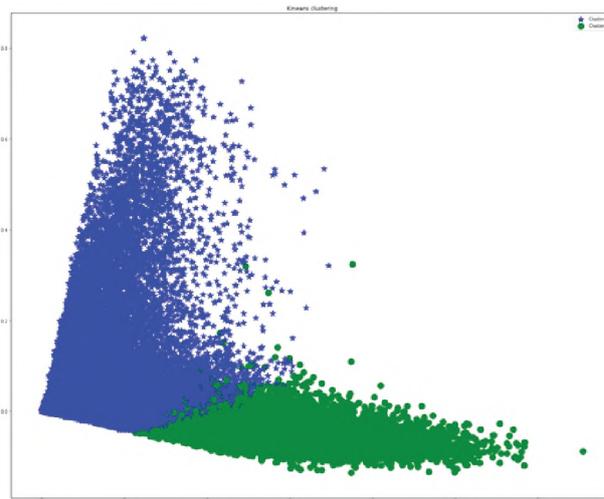


Рис. 5 – Графическое представление векторов tf-idf в 2D-пространстве, где зеленые метки – это депрессивные посты, а синие метки – обычные посты

**РЕЗУЛЬТАТЫ**

Если мы сравним результаты f1-score различных алгоритмов, мы увидим, что случайный лес с алгоритмом tf-idf показывает результат 95%, что является очень хорошим результатом.

И чтобы убедиться в правильности нашего алгоритма, был создан ROC с перекрестной проверкой [8].

«Крутизна» кривых ROC также важна, так как она идеально подходит для максимизации истинного положительного значения при минимизации ложного положительного значения.

На приведенном ниже графике показана кривая ROC для различных наборов данных о поездках и тестах, созданная в результате перекрестной проверки по К-средние критерию. Взяв все эти кривые, можно рассчитать среднюю площадь под кривой и увидеть дисперсию кривой, когда тренировочный набор разделен на разные подмножества. Это приблизительно показывает, как на выход классификатора влияют изменения в обучающих данных, и насколько отличаются расщепления, полученные при перекрестной проверке по К-кратному критерию, друг от друга.

## ВЫВОДЫ

В этой статье мы реализовали различные алгоритмы обучения с учителем и без учителя. Мы получили f1-score 95% и ROC-площадь 0,98 со случайным лесом с моделью векторизации tf-idf. По сравнению с нашим ранее построенным алгоритмом мы увеличили прогноз почти на 20%. Мы также проверили, как алгоритм обучения без учителя будет работать с этим набором данных, и она неожиданно показала отличные результаты.

Есть несколько интересных направлений будущей работы. Одна из них - реализовать модели глубокого обучения с помощью фреймворка PyTorch. Мы также создадим систему

оповещения для правительства, чтобы контролировать эмоциональное состояние человека, чтобы предотвратить возможные попытки самоубийства или какие-либо травмы, причиненные самому себе.

Мы подняли очень фундаментальный вопрос исследования об определении депрессивных постов в социальных сетях и обеспокоены анонимностью данных, особенно когда тема чувствительный. Мы контролировали параметры алгоритмов обучения и проверяли их кривой ROC, а также визуализированные результаты в 2D-пространстве. Также мы сделали это с открытым исходным кодом проект, для будущих изменений [9].

## ЛИТЕРАТУРА

1. World Health Organization. Preventing suicide. A resource for counsellors. Geneva 2006.
2. Oksana Lysenko. "The number of suicides among children in Kazakhstan continues is growing", [<http://www.zakon.kz/4524024-kolichestvo-suicidov-sredi-detejj-v.html>]
3. Klonsky E.D., May A.M. "Assessing Motivations for Suicide Attempts: Development and psychometric properties of the Inventory of Motivations for Suicide Attempts (IMSA)." *Suicide and Life-Threatening Behavior*, October 2013: 1-3.
4. Mukhtarkhanuly D., Abishev A. "Suicidal Post Detection in Social Networks using NLP." *Natural Sciences Publishing*, August 2018.
5. Marouane Birjali, Abderrahim Beni-Hssane, and Mohammed Erritali. "Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms." *International Arab Conference on Information Technology (ACIT'2016)*, 2016: 1-5.
6. Hardik A. Patel, Cheng-Yuan Hsieh - Knowledge Systems Institute. "Early Detection of Suicide Using Big-Data Analytics in Real Time." *Journal of Visual Languages and Sentient Systems*, 2016:1
7. Liu, Tong, Qijin Cheng, Christopher M. M. Homan, and Vincent M.B. Silenzio. "Learning from various labeling strategies for suicide-related messages on social media: An experimental study." *arXiv:1701.08796v1 [cs.LG]* 17, no. 01 (Jan 2017): 1-8.
8. Receiver Operating Characteristic with cross-validation [[http://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_roc\\_crossval.html#sphx-glr-auto-examples-model-selection-plot-roc-crossval-py](http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html#sphx-glr-auto-examples-model-selection-plot-roc-crossval-py)]
9. [https://github.com/DaniyarML/Publications/tree/master/Supervised\\_vs\\_Unsupervised](https://github.com/DaniyarML/Publications/tree/master/Supervised_vs_Unsupervised)

The ROC curve (Receiver Operator Characteristic) is the curve that is most often used to represent the results of binary classification in machine learning.