



Матеріали всеукраїнської науково-практичної конференції молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології (SoftTech-2022)»



2021



22-26 травня  
22-25 листопада  
Україна, Київ

## СКЛАД ПРОГРАМНОГО КОМІТЕТУ

**д.т.н., проф. Корнага Я.І.** – в.о. декана факультету інформатики та обчислювальної техніки;  
**д.т.н., проф. Жаріков Е.В.** – завідувач кафедри інформатики та програмної інженерії;  
**д.т.н., професор Снитюк В.Є.** – декан факультету інформаційних технологій, Київський національний університет імені Тараса Шевченка;  
**д.т.н., професор Купін А.І.** – завідувач кафедри комп'ютерних систем та мереж, Криворізький національний університет;  
**д.т.н., професор Чалий С.Ф.** – професор кафедри інформаційних управляючих систем, Харківський національний університет радіоелектроніки;  
**д.т.н., професор Гнатушенко В.В.** – професор кафедри інформаційних систем та технологій, Національна металургійна академія України;  
**д.т.н., професор Бабічев С.А.** – професор кафедри фізики та методики її навчання, Херсонський державний університет;  
**д.т.н., професор Литвиненко В.І.** – завідувач кафедри інформатики і комп'ютерних наук, Херсонський національний технічний університет;  
**д.т.н., професор Рудакова Г.В.** – професор кафедри автоматизації, робототехніки і мехатроніки, Херсонський національний технічний університет;  
**д.т.н., професор Павлов О.А.** – професор кафедри інформатики та програмної інженерії;  
**д.т.н., професор Стеценко І.В.** – професор кафедри інформатики та програмної інженерії;  
**д.т.н., професор Сидоров М.О.** – професор кафедри інформатики та програмної інженерії;  
**к.т.н., доцент Фіногенов О.Д.** – доцент кафедри інформатики та програмної інженерії;  
**к.т.н., доцент Лісовиченко О.І.** – доцент кафедри інформатики та програмної інженерії;  
**к.т.н., доцент Ліхоузова Т.А.** – доцент кафедри інформатики та програмної інженерії.

## СКЛАД ОРГАНІЗАЦІЙНОГО КОМІТЕТУ

**к.т.н., доцент Муха І.П.** – доцент кафедри інформатики та програмної інженерії;  
**к.т.н. Ліщук К.І.** – доцент кафедри інформатики та програмної інженерії;  
**к.т.н. Олійник Ю.О.** – доцент кафедри інформатики та програмної інженерії;  
**к.т.н., доцент Баклан І.В.** – доцент кафедри інформатики та програмної інженерії;  
**к.т.н., доцент Гавриленко О.В.** – доцент кафедри інформаційних систем та технологій;  
**к.е.н. Родіонов П.Ю.** – доцент кафедри інформатики та програмної інженерії;  
**Халус О.А.** – ст. викл. кафедри інформатики та програмної інженерії;  
**Лукутін О.В.** – ст. викл. кафедри інформатики та програмної інженерії;  
**Марченко О.І.** – ст. викл. кафедри інформатики та програмної інженерії;  
**Очеретяний О.К.** – асистент кафедри інформатики та програмної інженерії.

Збірник містить матеріали II та III Всеукраїнських науково-практичних конференцій молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2022), присвячених 125-й річниці КПІ ім. Ігоря Сікорського. Матеріали конференції. – Київ. – 2022. 22–26 травня та 23-25 листопада 2022 р. – 132 с.

У збірник включені тези доповідей, які були представлені на конференції «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2022). В доповідях розглянуті науково-практичні питання щодо сучасних аспектів інженерії програмного забезпечення і передових інформаційних технологій.

**Редакційна колегія:**

Баклан І.В., доцент, к.т.н, доцент кафедри ІІІ НТУУ «КПІ ім. Ігоря Сікорського»

Родіонов П.Ю., к.е.н, доцент кафедри ІІІ НТУУ «КПІ ім. Ігоря Сікорського»

Муравйова І.М., інженер I категорії кафедри ІІІ НТУУ «КПІ ім. Ігоря Сікорського»

Дизайн титульної сторінки: провідний інженер Майєр З.О. кафедри ІІІ НТУУ «КПІ ім. Ігоря Сікорського»

## ЗМІСТ

1	<i>КУЧМА АРТЕМ БОРИСОВИЧ ЖАРИКОВ ЕДУАРД В'ЯЧЕСЛАВОВИЧ</i>	ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ ПІДГОТОВКИ НАБОРУ ДАНИХ З КІЛЬКОХ ДЖЕРЕЛ ЗА ДОПОМОГОЮ ІНСТРУКЦІЙ ПРИРОДНОЮ МОВОЮ	5
2	<i>ПАВЛОВ ОЛЕКСАНДР АНАТОЛІЙОВИЧ ГОЛОВЧЕНКО МАКСИМ МИКОЛАЙОВИЧ ДРОЗД ВАЛЕРІЯ ВАЛЕРІЇВНА РЕВИЧ МАКСИМ МИКОЛАЙОВИЧ</i>	ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДУ ПОБУДОВИ БАГАТОВИМІРНОЇ ЛІНІЙНОЇ РЕГРЕСІЇ, ЗАДАНОЇ НАДЛИШКОВИМ ОПИСОМ	10
3	<i>АВСЕЦІН МАКСИМ ВІКТОРОВИЧ ЖАРИКОВ ЕДУАРД В'ЯЧЕСЛАВОВИЧ</i>	АКТУАЛЬНІ НАПРЯМКИ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ РОБОТИ ПРОГРАМНО-ВИЗНАЧЕНИХ МЕРЕЖ	14
4	<i>OLGA SOLOVEI</i>	A SELECTION OF DISTANCE METRIC FOR FEATURE SELECTION BY MUTUAL INFORMATION FILTER METHOD	19
5	<i>ПАЩЕНКО ВЛАДИСЛАВ ЮРІЙОВИЧ</i>	МЕТОД ЗГЛАДЖУВАННЯ ЗОБРАЖЕНЬ В СУЧАСНІЙ КОМП'ЮТЕРНІЙ ГРАФІЦІ	22
6	<i>ГОБОВ ДЕНИС АНДРІЙОВИЧ</i>	СУЧАСНІ ТРЕНДИ ПІДГОТОВКИ БІЗНЕС-АНАЛІТИЧНИХ ДОКУМЕНТІВ В ІТ-ПРОЕКТАХ	24
7	<i>VALERII NIKITIN EVGEN KRYLOV</i>	CONSISTENCY OPTIMIZATION METHODS IN DISTRIBUTED NOSQL DATABASES	27
8	<i>СМОЛЯР ГЕРМАН ВОЛОДИМИРОВИЧ ХАЛУС ОЛЕНА АНДРІЇВНА</i>	ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ АВТОМАТИЗАЦІЇ ТЕСТІВ	31
9	<i>СКРИГУН ВЛАДИСЛАВ ОЛЕКСАНДРОВИЧ</i>	МОДЕЛЮВАННЯ ЕКОСИСТЕМИ ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ	34
10	<i>ХІЛЬЧЕНКО ЄГОР АНДРІЙОВИЧ БАКЛАН ІГОР ВСЕВОЛОДОВИЧ</i>	РОЗРОБКИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ СТВОРЕННЯ ЛІНГВІСТИЧНИХ МОДЕЛЕЙ	37
11	<i>СУСЬКОВ ЯКІВ РОМАНОВИЧ</i>	РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ВИЗНАЧЕННЯ ФІЗИЧНОГО СТАНУ ЛЮДИНИ ЗА ДОПОМОГОЮ ПАРАМЕТРІВ МОВЛЕННЯ	39
12	<i>ЛИЦУК ІГОР СЕРГІЙОВИЧ</i>	ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ МЕДИЧНИХ СИСТЕМ ДІАГНОСТИКИ НА ОСНОВІ ТОМОГРАМ	42
13	<i>ШКУРКО ДЕНИС ОЛЕКСАНДРОВИЧ</i>	АВТОМАТИЧНЕ ВІДМІНЮВАННЯ АНТРОПОНІМІВ В ОФІЦІЙНИХ ДОКУМЕНТАХ	44
14	<i>ЧЕКОТУН ЯРОСЛАВ ДМИТРОВИЧ</i>	ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ РОЗПОДІЛЕНОЇ СИСТЕМИ ТОВАРООБІГУ	46
15	<i>КУВІЧКА МАКСИМ ЄВГЕНОВИЧ ОЛІЙНИК ЮРІЙ ОЛЕКСАНДРОВИЧ</i>	УНІФІКАЦІЯ СТРУКТУРИ НАДВЕЛИКИХ МАСИВІВ ТЕКСТОВИХ ДАНИХ ЗІБРАНИХ З РІЗНИХ ДЖЕРЕЛ	48
16	<i>ФЕДОРОВИЧ ІЛЛЯ АНДРІЙОВИЧ ОЛІЙНИК ЮРІЙ ОЛЕКСАНДРОВИЧ</i>	МОДЕЛІ ОБРОБКИ ПОТОКІВ ТЕКСТОВИХ ДАНИХ В РУШІІ АРАСНЕ SPARK STRUCTURED STREAMING	52
17	<i>ANDRII IHOROVYCH VASHCHENOK OLEH IVANOVYCH LISOVYCHENKO</i>	SOFTWARE ARCHITECTURE OF FAULT-TOLERANT BIG DATA PROCESSING SYSTEMS	57
18	<i>БОВСУНОВСЬКИЙ ОЛЕКСІЙ ЛЕОНІДОВИЧ ЛІХОУЗОВА ТЕТЯНА АНАТОЛІЇВНА</i>	ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ СЕРВІСУ КІНОФІЛЬМІВ	61

19	<i>СОКОЛОВСЬКИЙ ВЛАДИСЛАВ ВОЛОДИМИРОВИЧ ЖАРИКОВ ЕДУАРД В'ЯЧЕСЛАВОВИЧ</i>	АРХІТЕКТУРА ПРОГРАМНО АПАРАТНОЇ СИСТЕМИ МОНІТОРИНГУ СТАНУ ОБ'ЄКТІВ ПІДВИЩЕНОЇ НЕБЕЗПЕКИ З МОЖЛИВІСТЮ ПРОГНОЗУВАННЯ ВИНИКНЕННЯ НАДЗВИЧАЙНОЇ СИТУАЦІЇ	64
20	<i>НАГОРНИЙ МАКСИМ ЮРІЙОВИЧ ХАЛУС ОЛЕНА АНДРІЙВНА</i>	МЕТОД РОЗРОБКИ АТОМАРНИХ GRPC МІКРОСЕРВІСІВ	69
21	<i>ЩЕРБАКОВА ЮЛІЯ ОЛЕГІВНА ОЛІЙНИК ЮРІЙ ОЛЕКСАНДРОВИЧ</i>	МЕТОДИ РОЗПОДІЛЕНОГО НАВЧАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ ДІАГНОСТИКИ ЗАХВОРЮВАНЬ ЗА ДОПОМОГОЮ РЕНТГЕНІВСЬКИХ ЗНІМКІВ	72
22	<i>ПИЖ ОЛЕКСАНДР ІГОРОВИЧ КРАМАР ЮЛІЯ МИХАЙЛІВНА</i>	ІНТЕЛЕКТУАЛЬНА СИСТЕМА АНАЛІЗУ ТА КЛАСИФІКАЦІЇ ВХІДНИХ Е-МАІЛ ПОВІДОМЛЕНЬ НА ОСНОВІ ЇХ ВМІСТУ	76
23	<i>ШЕЛЕСТЮК МАКСИМ ІГОРОВИЧ ЖУРАКОВСЬКА ОКСАНА СЕРГІЙВНА</i>	ЗАДАЧА ФОРМУВАННЯ РОЗКЛАДУ В ІНФОРМАЦІЙНІЙ СИСТЕМІ ПІДТРИМКИ ДІЯЛЬНОСТІ СПОРТИВНОГО ЗАКЛАДУ	78
24	<i>САРНАЦЬКИЙ ВЛАДИСЛАВ ВІТАЛІЙОВИЧ БАКЛАН ІГОР ВСЕВОЛОДОВИЧ</i>	МОВНИЙ ЗАСІБ ОПИСУ АГЕНТНИХ ЕПІДЕМІОЛОГІЧНИХ МОДЕЛЕЙ	85
25	<i>ЖУРБА МИКОЛА АНДРІЙОВИЧ СТЕЦЕНКО ІННА В'ЯЧЕСЛАВІВНА</i>	ПОЛПШЕННЯ ПІДХОДУ СТВОРЕННЯ НОВИХ ЦІЛЕЙ ПРОЄКТУ XCODE З ВИКОРИСТАННЯМ СКРИПТІВ ДЛЯ КОДОГЕНЕРАЦІЇ	90
26	<i>ГАЛАЙКО ДЕНИС ОЛЕКСІЙОВИЧ ОЛІЙНИК ЮРІЙ ОЛЕКСАНДРОВИЧ</i>	SEARCHING TEXT SIMILARITY PARALLEL METHOD	94
27	<i>БОЙЧУН СОФІЯ ЄВГЕНІВНА ГАВРИЛЕНКО ОЛЕНА ВАЛЕРІЙВНА</i>	СИСТЕМА ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ ВИСТУПУ СПОРТСМЕНІВ	100
28	<i>СЕМЧЕНКО АНДРІЙ ОЛЕГОВИЧ ОЛІЙНИК ЮРІЙ ОЛЕКСАНДРОВИЧ</i>	МОНІТОРИНГ ПРОЦЕСУ РОЗРОБКИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ НА ОСНОВІ АНАЛІЗУ ТЕКСТОВИХ ДАНИХ	103
29	<i>ДОБРЯНСЬКИЙ БОГДАН ІГОРОВИЧ САВАСТРУ СТАНІСЛАВ ВІКТОРОВИЧ</i>	МОДЕЛЮВАННЯ ПРОГРАМНОЇ КОНФІГУРАЦІЇ ETL-ПРОЦЕСУ	111
30	<i>КОРЗУН ІЛЛЯ МИХАЙЛОВИЧ ПАВЛОВ ОЛЕКСАНДР АНАТОЛІЙОВИЧ</i>	МЕТОДИ І ПРОГРАМНІ ЗАСОБИ ВИЯВЛЕННЯ АНОМАЛІЙ ПРИ СКАНУВАННІ БАНКНОТ ОПТИЧНИМИ СЕНСОРАМИ	115
31	<i>ПАНАСЮК СТАНІСЛАВ ІВАНОВИЧ</i>	МЕТРИЧНИЙ ОПИС КОМПОНЕНТІВ БАГАТОКРАТНОГО ВИКОРИСТАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНИХ СИСТЕМ	121
32	<i>МИНЗАР БОГДАН МИКОЛАЙОВИЧ СТЕЦЕНКО ІННА В'ЯЧЕСЛАВІВНА</i>	ІНТЕЛЕКТУАЛЬНА СИСТЕМА ВИЯВЛЕННЯ ПРОПАГАНДИ В ТЕКСТАХ	123
33	<i>ПАЩЕНКО ІЛЛЯ МИКОЛАЙОВИЧ ЯЛАНЕЦЬКИЙ ВАЛЕРІЙ АНАТОЛІЙОВИЧ</i>	ЦЕНТРАЛІЗОВАНІ ТА ДЕЦЕНТРАЛІЗОВАНІ СИСТЕМИ ЕЛЕКТРОННОГО ГОЛОСУВАННЯ	127
34	<i>МИХАЙЛОВ ДАНИЛ ЄВГЕНОВИЧ НОВІНСЬКИЙ ВАЛЕРІЙ ПЕТРОВИЧ</i>	СПОСІБ ОПЕРАТИВНОЇ ІДЕНТИФІКАЦІЇ ВОЄННОЇ ТЕХНІКИ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ	129
35	ПРО АВТОРІВ		131

UDC 004.415.25

OLGA SOLOVEI

## A SELECTION OF DISTANCE METRIC FOR FEATURE SELECTION BY MUTUAL INFORMATION FILTER METHOD

A dimensionality reduction is almost a mandatory part of data pre-processing before to build a machine learning models. One of the ways to reduce a dimension of the dataset is to select features which are the most predict to target class. The current study is focused on Mutual Information method for feature selection. The goal of the work is to understand by conducting empirical studies if Mutual Information method that is implemented in Python's library sklearn with Chebyshev distance's metric has a negative impact on a classification model's performance.

**KEY WORDS:** Feature Selection, Mutual Information, distance metric, Area under ROC curve.

Зменшення розмірності є майже обов'язковою частиною попередньої обробки даних перед створенням моделей машинного навчання. Одним із способів зменшити розмірність набору даних є отбір ознак, які є найбільш доречними для прогнозування значення цільової змінної. Поточне дослідження зосереджено на методі взаємної інформації. Мета роботи полягає в тому, щоб шляхом проведення емпіричних досліджень зрозуміти, чи має негативний вплив на продуктивність моделі класифікації те, що метод взаємної інформації в бібліотеці sklearn мови програмування Python реалізованою з метрикою відстані Чебишева.

**КЛЮЧОВІ СЛОВА:** Feature Selection, Mutual Information, distance metric

**Introduction.** The goal of feature selection is to choose an optimal feature subset according to predefined evaluation criterion. Filter methods for feature selection are independent from any learning algorithm because their evaluation strategy is based on different statistical measures and thus are often faster and have a good generalization ability than other methods. The appropriate filter's method is selected depending on the parameters: a type of machine learning task; in case of a classification – a number of unique values in target class (binary or multi-class); type of predictive features – continuous/discrete or categorical; type of data in dataset – flat feature; structure feature; linked data; multi-source; multi-view; streaming feature; streaming data [1].

Mutual information (MI) as a filter method to select the most predictive features is often used because it is able to detect non-linear dependencies, which is its main strength, and it can be applied when dataset includes continues, discrete or both types of data [2,3]. MI has a straightforward interpretation as the amount of

shared information between independent feature and target class. MI in machine learning library sklearn is measured by k- nearest-neighbours based estimators, the method sometimes referred to as KSG (Kraskov-Stögbauer-Grassberger) [4]. Given MI between an independent feature  $X$  and a target class  $Y$  is  $I(X, Y)$  then according to KSG algorithm for each data point  $i$  the method computes a number  $I_i$  based on its nearest-neighbours in the continuous variable  $y$  as formalized per eq. 1 and MI is obtained as an average of  $I_i$  according to eq. 2.

$$I_i = \psi(N) - \psi(N_{x_i}) + \psi(k) - \psi(m_i) \quad (1)$$

where  $\psi(\cdot)$  is a digamma function [5];  $N$  – number of samples in dataset;  $N_{x_i}$  and  $m_i$  – are  $k$  - closest neighbours to point  $i$  among those whose values are within distance  $d$  which can be calculated using any selected distance's calculation method;  $k$  – value is selected manually.

$$I(X, Y) = \langle I \rangle_i = \psi(N) - \langle \psi(N_{x_i}) \rangle + \psi(k) - \langle \psi(m_i) \rangle \quad (2)$$

In Python library sklearn KSG method is implemented with distance  $d$  calculated by Chebyshev metric, however, others metrics, for example, Euclidean, normalized Euclidean, squared Euclidean, Correlation, Cosine, Bray-Curtis and others are left aside without possibility to choose them as method's parameter [6]. However, the selection of distance metrics can potentially impact the value of calculated MI and as a result the selected predictive features could be less correct compared to the results of the other filter methods. In current work, will be performed the experiments of selecting different metrics to define distance  $d$  while calculating MI by KSG method and evaluate whether a selection of the distance metric has a negative impact on a model's performance, when the model is created by Random Forest Classifier algorithm.

**Main part.** Our datasets are from UCI Machine Learning Repository: Pima Indians Diabetes ("pima diabetes"), E. Coli Genes ("coli") and Million Songs ("songs"). The features' count per each dataset is 9, 86 and 58 and samples' count

per each dataset is 768, 9822, 1409 correspondingly. In the experiments for each dataset are included the steps: 1) calculate MI using KSG method with distance metrics: Chebyshev, Bray-Curtis, Cosine, squared Euclidean, Correlation; 2) take features which MI's score is positive and not equal to 0; 3) train machine learner Random Forest Classifier with selected features and target class; 4) evaluate model's performance by area under ROC curve (AUC); precision and recall metrics.

The values of the performance metrics of Random Forest Classifier with features that are selected by chosen distance metrics are recorded in table 1. It is noticeable, that the highest value of AUC is when MI is applied with Correlation metric for "coli" dataset; with squared Euclidean distance method for "songs" dataset; with Cosine – "pima diabetes" dataset. The same is visible on ROC convex hull graphs - that "more northwest" ROC curve corresponds to distance metrics: Correlation, squared Euclidean, Cosine for each dataset correspondingly.

Table 1.

The performance metrics of Random Forest Classifier

Distance method	"coli"			"songs"			"pima diabetes"		
	AUC, %	Precision, %	Recall, %	AUC, %	Precision, %	Recall, %	AUC, %	Precision, %	Recall, %
Chebyshev	64	52	51	46	53	52	81	66	61
Bray-Curtis	56	51	50	44	55	54	79	61	62
Cosine	55	50	50	43	56	54	82	67	65
Squared Euclidean	67	52	51	47	54	53	81	65	67
Correlation	68	52	51	44	55	54	80	62	62

The values of precision metric from table 1 show that the ability of Random Forest classifier not to label as positive a sample that is negative is higher when AUC has the highest

value. Similar tendency is visible for Recall, i.e., the ability of Random Forest classifier to find correctly all positive samples had increased when AUC had the highest value.

### Conclusions.

Based on the results from conducted experiments can be concluded that the selection of the distance metric has an impact on classification's model performance. Enabling the ability to select the distance metric while calculating a mutual information by algorithm from Python's library sklearn could help to improve a classification's precision and recall results simultaneously.

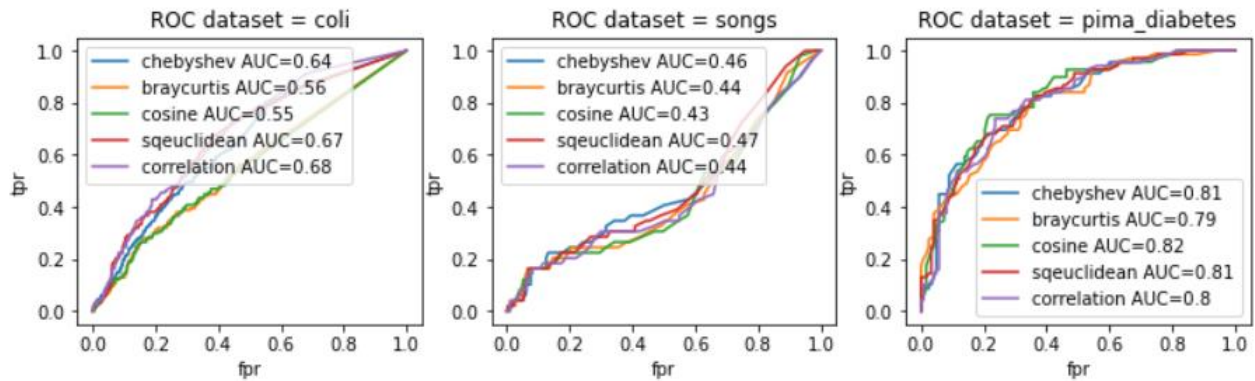


Fig. 1. ROC convex hull of Random Forest Classifier

### References

1. Feature selection: A data perspective/ Jundong Li, et al // ACM computing surveys (CSUR).-2012.-Vo.50(6).-pp.1-45.
2. Estimating mutual information for discrete-continuous mixtures/Gao, Weihao, et al // Advances in neural information processing systems 30.-2017.
3. Mutual information between discrete and continuous data sets/Ross B. C.// PloS one.-2014.-Vo.9(2).-p.e87357.
4. Estimating mutual information/Kraskov, A., Stögbauer, H. and Grassberger, P.// Physical review E.-2004.-Vo.69(6).-p.066138.
5. Abramowitz, Milton, Irene A. Stegun, and Robert H. Romer. "Handbook of mathematical functions with formulas, graphs, and mathematical tables." (1988): 958-958.
6. sklearn.feature\_selection.mutual\_info\_classif [Электронный ресурс].