DOI: 10.31489/2023Ec1/59-66
JEL I20
UDC 338.246.028

D.K. Ilyassov[1*] K. Kitapova[2] T. Kenc[3]

[1]Narxoz University, Kazakhstan
[2]Almaty Technological University, Kazakhstan
[3]INCEIF, Malaysia

[1]didar.ilyassov@narxoz.kz
[2]kulzira.tsagaankhuu@mail.ru [3]turalay.kenc@gmail.com)

[1]Scopus Author ID: 0000-0001-6150-6492
[2] Scopus Author ID 0000-0002-8479-173X
[3] Scopus Author ID 0000-0001-5051-3726)

Overview and advantages of Machine Learning (ML) in Statistics

Abstract

Object: The main purpose of this study is to provide insight into why machine learning is the future of statistics. The virtual world generated a vast amount of data bringing together intelligent machines and networked processes. Machine learning as the emerging field of data science leads to new implications for statistics in terms of the big data era. Nowadays Machine Learning (ML) application is becoming broader including psychology, artificial intelligence, control theory, information theory, neuroscience, philosophy, Bayesian method, computational complexity theory etc. The recent use of ML in medicine, agriculture or trading is evidence of its future development in the coming years.

Methods: This study is based on the literature review of Machine learning (ML) models, paradigms, algorithms, and their advantages versa classical statistics. As obvious of ML application, the number of articles on Machine Learning and Data Science vs Classical Statistics in Wikipedia reflected in Python.

Findings: The main results of this study are listing the main Machine Learning Algorithms and applications. In addition, this paper identifies the main advantages and disadvantages of Machine Learning versa classical statistics.

Conclusions: There are many advantages of Machine Learning (ML), which highlight the future of Machine learning methods in statistics. The increase in data and innovations make a long and broad way of Machine Learning (ML) development.

Keywords: Machine Learning, Supervised Learning, Unsupervised Learning, Reinforcement Learning, Evolutionary Learning, Semi-Supervised Learning, Neural Network, Data Science

Introduction

Today Machine Learning (ML) can be applied in various directions of psychology, artificial intelligence, control theory, information theory, neuroscience, philosophy, Bayesian method, computational complexity theory etc.

Machine Learning might solve mostly five different problems. The first classification problem answers the question “Is this A or B?” Anomaly detection problem occurs to identify the odd one to make out. How many quantitative questions are related to the regression problem? The organizing and hidden issues behind a problem are called as a clustering problem. The reinforcement problem is devoted to anticipating the next things that will happen.

The development of ML starts in the 1950s when introduced Turning Test persuade people that they talked with humans, not with machines. The last social network developments lead to innovations such as Deep Learning, Amazon, and Google platforms.

The virtual world generated vast amounts of data bringing together intelligent machines and networked processes. Machine learning as the emerging field of data science leads to new implications for statistics in terms of the big data era. Nowadays Machine Learning (ML) application is becoming broader including psychology, artificial intelligence, control theory, information theory, neuroscience, philosophy, Bayesian method, computational complexity theory etc. The recent use of ML in medicine, agriculture, or trading is evidence of its future development in the coming years.

It has actually to compare Machine Learning Algorithms with classical statistics by showing the pros and cons. This study is making comparisons and finds key ideas through a literature review. Therefore, based on secondary data the advantages and drawbacks of ML are provided. Moreover, this analysis highlights the future growth and development opportunities of Machine Learning (ML) in the coming years.

* Corresponding author’s e-mail: *didar.ilyassov@narxoz.kz*

Literature Review

Similarly, to Big Data Analytics the theoretical framework of Machine Learning (ML) is building up. It is obvious that new technologies will bring new methods of ML. However, the six algorithm steps create a machine-learning model similar to other data processing. The steps, tasks and brief description is given in Table 1. Today Machine Learning (ML) can be applied in various directions of psychology, artificial intelligence, control theory, information theory, neuroscience, philosophy, Bayesian method, computational complexity theory etc.

Table 1. The components of ML with the main tasks and brief description

Steps	The main task	Brief description
Step I – Data Set collection and preparation	To format data as input to the algorithm	The cleaning of noise or irrelevant data to make it to a structured format.
Step II – Feature selection	To remove irrelevant features	The selection of the most important features subset.
Step III – Algorithm selection	To choose the most suited algorithm for problem solution	There are many various learning algorithms. The most imperative for the best possible results should be applied.
Step IV – Model and Parameters choice	To set the most appropriate parameter values of algorithms	Some initial manual invention helps to identify the most suitable model and parameters.
Step V – Exercising	To train model using a part of data set	The use of training data to improve model application.
Step VI – Performance assessment	To assess model application by using accuracy, precision and recall performance parameters	The model testing before real-time application to confront unobserved data how it meets performance parameters.
Note: adopted by authors from Alzubi, et al., 2018 and Batta, 2019		

Machine Learning recently used in different fields as shown in Figure 1.

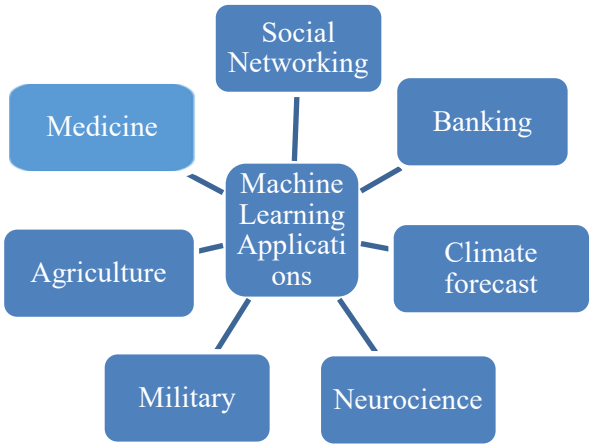

Figure 1. Machine Learning (ML) applications

Note: adopted by authors from Alzubi, et al., 2018 & Khan, A., 2010

The paradiйms of ML with the main tasks, algorithms, and brief descriptions are provided in Table 2. In accordance with Table 2, the algorithm's training and output availability classify ten categories of Machine Learning (ML) paradigms. According to Alzubi, et al., 2018 among mentioned paradigms in Table 2 supervised learning is staying as the most popular.

Table 2. The paradigms of ML with the main tasks, algorithms and brief description

Learning paradigm	The main task	ML popular Algorithms	Brief description	ML application and examples
Supervised Learning	To make applications from predictions of historical data	Decision Tree Naïve Bayes Support Vector Machine Regression Analysis	The outputs in the case of classification are discrete and continuous for regression	Recognition systems and e-commerce website offerings. Classification and regression.
Unsupervised Learning:	To find some features, patterns and rules from the data	Principal Component Analysis K-Means Clustering	Learning and revealing some structure in unlabeled data	In the case of unknown data categories are suitable Feature vectors to apply predictive models for text, documents, images, etc. Clustering, association, dimensionality reduction
Reinforcement learning	To get the correct output	As learning to check the output correctness	Learning	No problem solving, but applied in classification and control
Evolutionary Learning	To adapt inputs and rules by behavior	To propose the best solution to the problem	Understanding by algorithm	Mostly applied for biological organisms to adapt their environment
Semi-Supervised Learning	To use the power of supervised and unsupervised learning	Generative Models. Self Training. Transductive Support Vector Machine.	Best suitable to model building by a lack of skills and high cost of observations	Generative Models, Self-Training and Transductive SVM are generated categories to use. It can be used for problems like classification, regression and prediction
Ensemble Learning	To set many hypotheses to build a prediction model	Random Forest	Bias decreasing (boosting), Variance (bagging) and precise predictions (stacking) Random Forest as parallel explosion of relationships among base learners	AdaBoost tests the dependence between the common learners Boosting reflects the sequence of weak models in a small number of observations AdaBoost is adaptive boosting. Bagging as bootstrap gets means of all predictions.
Neural Network	To adjust the weights of neuron as a nerve cell interconnections by help of electric impulses to distribute through the brain	Supervised Neural Network Unsupervised Neural Network Unsupervised Neural Network Reinforced Neural Network	Adjusting weights help to get accurate results by training to make predictions on unobserved data. Also make group them by similarities to get correct outputs	Training data and data classification by similarities or human learning by mistakes in interaction with the environment based on past decisions.
Instance based learning	To generalize based on training data	K-Nearest Neighbor, k-means, k-medians, hierarchical clustering and expectation maximization	Any inputs can be compared with the trained instances to make predictions	Database of training instances allows to apply differently, globally and locally in an easy way quickly with some time for prediction
Dimensionality reduction algorithms	To deal with high dimensionality and sparsity of data to make implicit data structure	Multidimensional scaling (MDS), Principal component Analysis (PCA), Linear Discriminant Analysis (LDA), Principal component regression (PCR), and Linear Discriminant Analysis (LDA)	Reducing dimensions help to avoid irrelevant and redundant data to get higher accuracy of results	Applications in climatology, biology, astronomy, medical, economy and finance
Hybrid Learning	To decrease errors of ensembles by hybridization to make heterogeneous models	Heterogeneous models by combining clustering with association mining or decision tree etc.	In classification algorithms to decrease of computational complexity, over fitting and sticking to local minima by model combinations	Solving complex tasks with error minimization

Note: adopted by authors from Alzubi, et al., 2018

Machine learning (ML) applications are broadly used in different fields of life: computer games, sophisticated speech recognition systems, driving autonomous vehicles, filter spam emails, robotics and artificial intelligence, text mining, emotion reflections, document categorization, search engines, web marketing, text classification in social networking, medical field, banking, facial recognition climate forecast, stock trading systems (Khan, 2010).

Prediction of the moon cycle, seasons, and future agriculture yields, humankind is getting information from indirect observations and needs future intersections of statistics and data science (Sofia et al., 2019). However, this study explores Machine learning Algorithms and applications.

Methods

This study is based on a literature review of Machine learning (ML) models, paradigms, algorithms, and their advantages versa classical statistics. As obvious of ML application, the number of articles on Machine Learning and Data Science vs Classical Statistics in Wikipedia reflected in Python.

The recent trends of Machine Learning and Data Science development from 2016 to 2022 as mentioned above are implemented by using codes of Python. Such visualization of big data also highlights the advantages of Machine Learning.

Results

The application of Machine Learning faces many challenges. As mentioned by Alzubi, et al., 2018 they are:

- Machine learning methods require a big amount of data to make accurate results and predictions. However, researchers are not always able to get such an amount of data. In this case, giants like Facebook and Google are leading in the field of Artificial Intelligence.
- Spam detection. It is not easy still to detect spam or not.
- Machine learning algorithms still have problems in differentiating objects and images. Deep learning algorithms and different fields of Machine Learning use are new challenges.

Machine Learning Algorithms are continuously developing widespread spreading everywhere (Bhatia & Kumar, 2017). Today the following applications can be highlighted: deep learning, data mining, and data analytics, natural language processing, testing and simulation, machine learning in medicine, and human-computer interactions (Christian et al., 2021, Sarker, I., 2021, Xuming et al., 2020). Machine Learning (ML) Wikipedia page views from 2016 to 2022 is illustrated in Figure 2 that shows the popularity of it. The peak of popularity was in 2019 that then slowed down by COVID -19. Below provided codes of Python for Machine Learning that help to collect and analyze actual data (Fig. 2).

```
p = PageviewsClient(user_agent="Python query script by " + your_contact
_info)
MLviews = p.article_views(project='en.wikipedia', articles=['Machine
Learning', 'Artificial Intelligence'], granularity='monthly',
start='20160101', end='20221231')
ML_df = pd.DataFrame(MLviews)
ML_df = ML_df.transpose()
ML_df = ML_df.set_index(ML_df.index.strftime("%Y-%m")).sort_index()
ML_df
fig = plt.figure()
plt.title("Monthly Wikipedia pageviews for ML")
plt.ticklabel_format(style = 'plain')
ax = ML_df.iloc[:,0].plot(kind='line', figsize=[14,8], color="purple")
ax.set_xlabel("Monthly pageviews")
ax.set_ylabel("Month")Text(0, 0.5, 'Month')
```

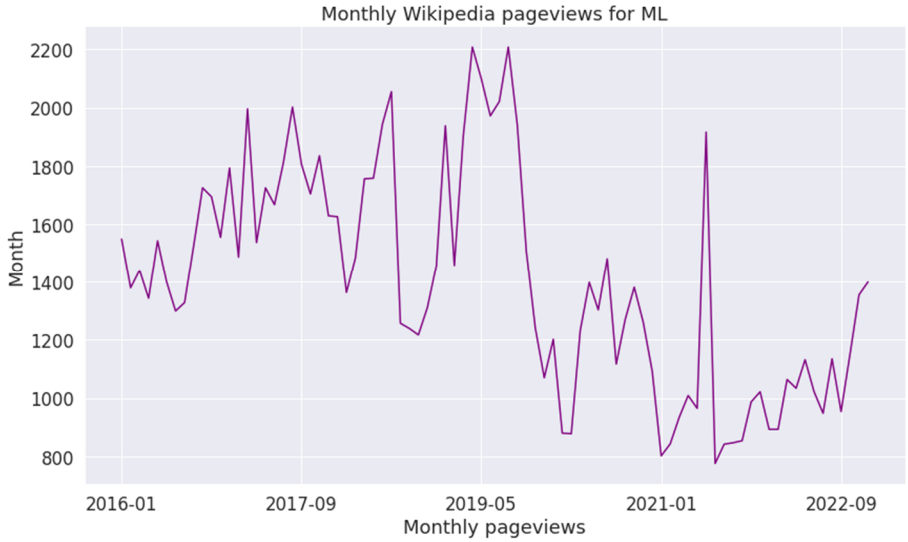

Figure 2. Machine Learning (ML) Wikipedia page views from 2016 to 2022

Note: moderated by authors by using Python

Nowadays SMAC (Social, Mobile, Analytic, and Cloud) technology expand the borders of ML application due to the big data rise as well. Computers are becoming powerful by ML algorithms and taking human-like behavior. Digitalization of any activities makes outputs precise and fast coming. The digitalization of government services also highlights the importance of ML applications (Kumar et al., 2017). According to our analysis in Figure 3 it is obvious the popularity of Data Science versa Classical Statistics in Wikipedia.

```
fig = make_subplots(specs=[[{"secondary_y": True}]])
# Add traces
fig.add_trace(go.Scatter(x=df.index,y=df['Statistics'], name="Statistics"),
   secondary_y=False,)
fig.add_trace(go.Scatter(x=df.index,y=df['Data_Science'],
   name='Data Science',
   line=dict(color='red'),
   mode='lines'), secondary_y=True)
fig.update_layout(title='Data Science vs Statistics',
                  xaxis_title='Year')
fig.show()
```

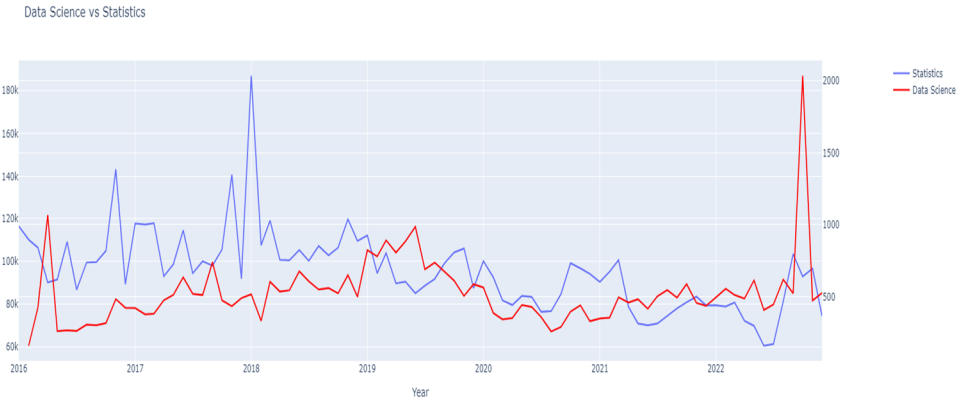

Figure 3. Data Science versa Classical Statistics

Note: moderated by authors by using Python

As evidenced by Figure 3 Data science is becoming more popular than Statistics. The increase in popularity of Data science including Machine Learning (ML) by innovations is obvious in the future.

Discussions

Data science including Machine Learning (ML) is developing rapidly than classical statistics. Table 3 shows the main categories of Big Data Analytics with its main purpose, advantages, and disadvantages versa classical statistics.

Table 3. Category of Big Data Analytics

Feature	Programming Language	Main purpose	Advantage	Disadvantage
WEKA	Java	Supervised and unsupervised data mining	Read files from numerous different database	Does not support much visualization
Rapid Miner	Java	Supervised and unsupervised data mining	Offers numerous procedures for selection of attribute and outlier detection	Consumes lots of RAM user computer, a large amount of data can produce an error
Orange	Python	Supervised and unsupervised data mining	Used for data visualization with mining technique	Working with a limited scale of data, additional libraries need to download
Tableau	No	Visualization	Low cost, less capacity of memory and easy to upgrade	Not support statistical features and need to integrate with other software platforms
R programming	C++, Fortran, R	Supervised and unsupervised data mining	No restriction for R license and compatible across platforms	Lack of memory management that caused by any available memory when needed quick task peforms
KNIME	Java	Supervised and unsupervised data mining	Capability to process massive data that only can be limited on the available computer hard disk space	Update to the latest version not working unless user installing the software again

Note: adopted by authors from Nor et al, 2020

As the commonly used Machine Learning Methods (ML) Tree-based methods observe inputs and the responses assuming data generating as complex and unknown (Shafiee et al., 2020). It interacts with variables by revealing some hidden patterns. Other algorithms allow learning from the Data as well (Masci et al., 2017). It manages to find complex and very flexible functional forms in the data without simply over fitting (Mullainathan et al., 2017). Boosting, bagging and random forests serve to reduce variance and increase predictive power (James et al., 2013). CART Model is able to find interaction to fit non-linear relationships over individual CARTs (Friedman, 2001).

As discussed Big Data Analytics including Machine Learning (ML) allow new opportunities and challenges.

Conclusions

By comparison of Classical Statistics and Machine Learning, the main advantages of Machine Learning are:

- Learn from the Data (Al-Jarrah et al., 2015).
- Tree-based methods can be classified as Machine Learning Methods (ML). It observes inputs and the responses assuming data generating is complex and unknown. It interacts with variables by revealing some hidden patterns (Masci et al., 2017).
- It manages to find complex and very flexible functional forms in the data without simply over fitting (Mullainathan et al., 2017).

- Boosting, bagging, and random forests serve to reduce variance and increase predictive power (James et al., 2013).

- CART Model is able to find interaction to fit non-linear relationships over individual CARTs (Friedman, 2001).

A lesser amount of data is suitable for supervised Learning, and better performance and results might be obtained by Unsupervised Learning with big data, but when data is becoming huge it is better to apply deep learning (Batta, 2019).

To sum up, the future of Machine learning methods in statistics is clear in a long and wide way.

References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87-93.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: an overview. *Journal of physics: conference series, 1142*(1), 012012. IOP Publishing.Doi: 10.1088/1742-6596/1142/1/012012.

Batta, M. (2019). Machine Learning Algorithms – A Review. *International Journal of Science and Research,* 381-386. Doi:10.21275/ART20203995.

Bhatia, M., & Kumar, A. (2017). Information Retrieval &Machine Learning: Supporting Technologies for Web Mining Research & Practice, *Webology, 5, 2,* 1-3. Retrieved from https://www.webology.org/abstract.php?id=106.

Christian, J., Patrick, Z., & Kai, H. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-95.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning, 112* (18). New York: springer.

Khan, A., Baharudin B., & Lan, H. (2010). A Review of Machine Learning Algorithms for Text Documents Classification. *Journal of Advances in Information Technology, 1*(1), 1-8. Doi: 10.4304/jait.1.1.4-20.

Kumar, A., & Sharma, A. (2017). "Systematic Literature Review on Opinion Mining of Big Data for Government Intelligence". *Webology, 14(*2), 1-8. Retrieved from ttp://www.webology.org/2017/v14n2/a156.pdf.

Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072-1085. Doi: 10.1016/j.ejor.2018.02.031.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106. Doi:10.1257/jep.31.2.87.

Nor, S. & Mutalib, S. (2020). Prediction of Mental Health Problems among Higher Education Student Using Machine Learning. *International Journal of Education and management Engineering*, *10(6),* 1-9. Doi: 10.5815/ijeme.2020.06.01.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput Sci 2,* 160. Doi: 10.1007/s42979-021-00592-x.

Shafiee, N. S. M., & Mutalib, S. (2020). Prediction of mental health problems among higher education student using machine learning. *International Journal of Education and Management Engineering (IJEME)*, *10*(6), 1-9. Doi: 10.5815/ijeme.2020.06.01.

Sofia, O. & Patrick, W. (2019). The future statistics and data science. *Statistica and Probability Letters, 136,* 46-50. Doi: 10.1016/j.spl.2018.02.042.

Xuming, H., & Xihong, L. (2020). Challenges and Opportunities in Statistics and Data Science: Ten Research Areas, *Harvard Data Science Review, 1-8.* Doi: 10.1162/99608f92.95388fcb.

Д.К. Ильясов, К. Китапова, Т. Кенч

Статистикадағы машиналық оқытуға шолу және артықшылықтары

Аңдатпа:

Мақсаты: Зерттеудің негізгі мақсаты — машиналық оқыту статистиканың болашағы екендігі туралы түсінік беру. Виртуалды әлем ақылды машиналар мен желілік процестерді біріктіру арқылы көптеген деректерді жасайды. Машиналық оқыту деректер ғылымының дамып келе жатқан саласы ретінде үлкен деректер дәуірі тұрғысынан статистика үшін жаңа салдарға әкеледі. Қазіргі уақытта машиналық оқытуды (МО) қолдану кеңейіп келеді, оның ішінде психология, жасанды интеллект, басқару теориясы, ақпарат теориясы, неврология, философия, Байес әдісі, есептеу күрделілігі теориясы және т.б. МО-ны медицинада, ауыл шаруашылығында немесе саудада жақында қолдану оның алдағы жылдарда одан әрі дамуын көрсетеді.

Әдісі: Бұл зерттеу классикалық статистикамен салыстырғанда машиналық оқыту модельдері (МО), парадигмалар, алгоритмдер және олардың артықшылықтары туралы әдебиеттерді шолуға негізделген. МО қолданудан көріп отырғанымыздай, Википедиядағы классикалық статистикамен салыстырғанда машиналық оқыту және деректер туралы мақалалар саны Python-да көрсетілген.

Қорытынды: Бұл зерттеудің негізгі нәтижелері машиналық оқытудың негізгі алгоритмдері мен қосымшалары. Сонымен қатар мақалада классикалық статистикамен салыстырғанда машиналық оқытудың негізгі артықшылықтары мен кемшіліктері анықталған.

Тұжырымдама: Статистикадағы машиналық оқыту әдістерінің болашағын көрсететін машиналық оқытудың (МО) көптеген артықшылықтары бар. Деректер мен инновациялар көлемінің артуы машиналық оқытуды (МО) дамытудың ұзақ және кең жолын құрайды.

Кілт сөздер: машиналық оқыту, мұғаліммен оқыту, мұғалімсіз оқыту, күшейтілген оқыту, эволюциялық оқыту, мұғаліммен аралас оқыту, нейрондық желі, деректер ғылымы.

Д.К. Ильясов, К. Китапова, Т. Кенч

Обзор и преимущества машинного обучения в статистике

Аннотация

Цель: Основная цель этого исследования — дать представление о том, почему машинное обучение (МО) — это будущее статистики. Виртуальный мир генерирует огромное количество данных, объединяя интеллектуальные машины и сетевые процессы. Машинное обучение, как развивающаяся область науки о данных, приводит к новым последствиям для статистики с точки зрения эпохи больших данных. В настоящее время применение машинного обучения становится все шире, включая психологию, искусственный интеллект, теорию управления, теорию информации, неврологию, философию, байесовский метод, теорию сложности вычислений и т. д. Недавнее использование МО в медицине, сельском хозяйстве или торговле свидетельствует о его дальнейшем развитии в ближайшие годы.

Методы: Это исследование основано на обзоре литературы по моделям машинного обучения (ML), парадигмам, алгоритмам и их преимуществам по сравнению с классической статистикой. Как видно из применения ML, количество статей о машинном обучении и науке о данных, в сравнении с классической статистикой в Википедии, отражено в Python.

Результаты: Основными результатами этого исследования являются алгоритмы и приложения машинного обучения. Кроме того, авторами определены основные преимущества и недостатки машинного обучения по сравнению с классической статистикой.

Выводы: Есть много преимуществ машинного обучения, которые подчеркивают будущее методов машинного обучения в статистике. Увеличение объема данных и инноваций прокладывает долгий и широкий путь развития машинного обучения.

Ключевые слова: машинное обучение, обучение с учителем, обучение без учителя, обучение с подкреплением, эволюционное обучение, обучение смешанное с учителем, нейронная сеть, наука данных.